

Ig3D: Integrating 3D Face Representations in Facial Expression Inference

Lu Dong ^{*} , Xiao Wang ^{*} , Srirangaraj Setlur , Venu Govindaraju , and Ifeoma Nwogu 

State University of New York at Buffalo
<https://dongludeeplearning.github.io/Ig3D.html>

Abstract. Reconstructing 3D faces with facial geometry from single images has allowed for major advances in animation, generative models, and virtual reality. However, this ability to represent faces with their 3D features is not as fully explored by the facial expression inference (FEI) community. This study therefore aims to investigate the impacts of integrating such 3D representations into the FEI task, specifically for facial expression classification and face-based valence-arousal (VA) estimation. To accomplish this, we first assess the performance of two 3D face representations (both based on the 3D morphable model, FLAME) for the FEI tasks. We further explore two fusion architectures, intermediate fusion and late fusion, for integrating the 3D face representations with existing 2D inference frameworks. To evaluate our proposed architecture, we extract the corresponding 3D representations and perform extensive tests on the AffectNet and RAF-DB datasets. Our experimental results demonstrate that our proposed method outperforms the state-of-the-art AffectNet VA estimation and RAF-DB classification tasks. Moreover, our method can act as a complement to other existing methods to boost performance in many emotion inference tasks.

Keywords: 3D Face Representations · Facial Expression Inference · Intermediate and Late Fusion

1 Introduction

Facial expressions play a significant role in social interactions, as it can provide insights into a person’s feelings toward other individuals or events. Mehrabian and Wiener [46] suggest that 55% of communication is perceived through facial expressions. AI-based automated emotion analysis enhances user experience and it has well-known applications in autonomous driving [9], online course learning [4], security [42, 49], healthcare [48], medical rehabilitation [68], employee job retention [52], and many other social situations.

^{*} Equal contribution.

In the domain of facial expression inference (FEI)¹, there are two primary tasks: discrete and continuous facial expression inference. Discrete or categorical inference aims to assign facial expressions into distinct emotional categories, such as anger, sadness, joy/happiness, surprise, fear, disgust, contempt and neutral, whereas continuous inference assigns facial expressions within a continuous 2D numerical space; where the two dimensions are extents of *valence* (the extent of pleasant or unpleasant response to an emotional stimulus) and *arousal* (the state of physiological activation and alertness resulting from an emotional stimulus). Discrete classification has advanced quickly due to the ease of data annotation. However, the highly abstract nature of labels makes cross-cultural consensus difficult [2]. Additionally, experts disagree on the number of emotion categories [13, 14, 19, 55], which have increased from 7 to 135 [7]. Continuous circumplex emotion modeling [56] further quantifies emotions and promotes research, but it is challenging to annotate accurately and suffers from bias [32, 47]. To mitigate this, researchers typically try to increase the number of human observers to reduce bias [53]. Although both types of analysis have certain data limitations, combining them offers complementary benefits. Therefore, we believe that the model's analysis should consider both perspectives.

Recently, 3D mesh reconstruction from static human-centric images has shown impressive achievements across various applications [12, 57, 66, 70]. In particular, reconstructing 3D faces from monocular images using facial geometry has proven effective in capturing extreme, asymmetric, and subtle expressions accurately. Regressing parameters from images as a lightweight 3D representation can disentangle facial shape and expression, and easily generate 3D facial geometry using a morphable model like FLAME [40]. Therefore, a natural progression is to integrate this technology into facial expression inference. However, limited research has investigated the significance and impact of the parameters involved, and how they can enhance FEI. In this work, we investigate the performance of two latest 3D face regression models, EMOCA [10] and SMIRK [54], in the context of the FEI task.

Data fusion is a process dealing with data and information from multiple sources to achieve improved information for decision-making [17]. This paper addresses the fusion of information from 2D images with parameters regressed from 3D perspectives, presenting significant challenges and uncertainties. A major issue is the heterogeneity of feature representations. Features from different modalities can vary greatly, requiring effective methods to integrate these diverse features seamlessly. Therefore, this paper proposes two architectures, intermediate fusion and late fusion, to investigate the impact of different fusion methods on facial expression inference performance.

In summary, the contributions of this work are as follows:

First, We provide insights into the parameters of 3D face representation (pose, shape, expression, jaw, etc.) and compare the two recent 3D face

¹ Wagner *et al.* [62] postulate that facial expressions are inferred and not recognized, hence the use of the term FEI, rather than the more popular term FER - facial expression recognition.

representation models, SMIRK and EMOCA on the benchmark datasets. Experiments showcase that EMOCA 3D representation achieves better performance on FEI tasks.

We introduced two architectures for integrating 3D representations: intermediate fusion and late fusion. Our experiments demonstrate that the late fusion architecture achieves superior performance.

Lastly, we present a simple and effective architecture that can be flexibly adapted to various affective reasoning tasks. Extensive experiments demonstrate the efficiency of our method, with results surpassing the state-of-the-art in AffectNet VA estimation and RAF-DB classification. We will release the 3D representation data along with the code developed for this work.

2 Related Work

Facial Expression Inference Datasets

The continuous development of datasets has driven the advancement of AI-based affective models. Early datasets such as JAFFE [45], CK+ [44], and KDEF [6] primarily collected expressions under 7 or 8 discrete categories. Subsequently, in-the-wild datasets like FER2013 [16], AffectNet [47], RAF-DB [38,39], and Aff-Wild2 [22–34, 69] further increased the scale of data. EmotionNet [3], LIRIS_CSE [21] and MAFW [43] extended the discrete labels by including compound annotations to better capture the richness of human emotions. Additionally, fine-grained discrete labels have been explored in datasets like F2ED [72] and Emo135 [63]. Having both discrete and continuous annotations also opens up new avenues for research breadth [1,35,47] Contextual information is consciously considered and collected in the EMOTIC [35] and VEATIC [53] datasets.

To better compare performance advantages, this paper selects the widely studied AffectNet and RAF-DB datasets for analysis. AffectNet includes 287,651 training images and 3,999 validation images (typically used as test data in experiments). It provides annotations for 8 distinct categories and VA (Valence-Arousal) annotations. RAF-DB includes 12,272 images for training and 3,096 images for testing, covering 7 distinct categories. Additionally, the dataset contains labels related to attributes such as gender, age, and more.

Facial Expression Inference Models

Discrete expression inference on datasets like AffectNet and RAF-DB has made continuous progress. The current top accuracy models [51] on both datasets are the same: DDAMFN, FMAE, BTN, and ARBEx. DDAMFN [71] integrates a Mixed Feature Network (MFN) as the backbone and a Dual-Direction Attention Network (DDAN) as the head. FMAE [50] introduces Identity Adversarial Training (IAT) and pre-trains a Facial Mask Autoencoder. S2D [8] proposes the Static-to-Dynamic Model to improve Dynamic Facial Expression Recognition

(DFER) in videos. BTN [18] includes Multi-Level Attention (MLA) and Batch Transformer (BT) modules to address uncertainty and noisy data in Facial Expression Recognition (FER). ARBEx [64] incorporates learnable anchors and a multi-head self-attention mechanism in the embedding space to tackle class imbalance, bias, and uncertainty in expression learning tasks.

Regarding continuous expression VA inference, CAGE [62] is the current state-of-the-art framework. It uses a small-scale pre-trained version of the Multi-Axis Vision Transformer (MaxViT) [61] along with a lightweight EfficientNet model [59]. The core insight is to train both discrete category and VA estimation simultaneously, using the combined checkpoint to enhance VA inference. Therefore, DDMFN and CAGE frameworks are employed as our 2D image-side backbones for fusion analysis.

Data Fusion

Intermediate fusion allows data fusion at different stages of model training by transforming input data into higher-level feature representations through multiple layers [37]. It offers flexibility at different depths of fusion. In the context of deep learning with multimodal data, intermediate fusion involves merging different modal representations in a single hidden layer, enabling the model to learn joint representations. To improve performance, data dimensionality can be adjusted [11, 67].

Late fusion involves independently processing data sources at the decision stage before fusing the results [37]. This technique is inspired by the popularity of ensemble classifiers [36]. When data sources differ significantly in sampling rate, data dimensionality, and measurement units, this technique is much simpler than early fusion methods. Since errors from multiple models are handled independently, they are uncorrelated, and late fusion typically yields better performance. Many researchers use late or decision-level fusion to address multimodal data problems [20, 58, 65].

3 Face Representation with 3D Morphable Models

Although there are many different 3D face representation models, in this work, we are mainly considering 3D morphable models (3DMM), which have been widely used in widely used in facial geometry reconstruction tasks. These models regress the 3D representations of faces from 2D images, by projecting the input 2D image onto the previously established 3DMM space, to obtain shape, pose, expression, and detail coefficients. Nonlinear optimization techniques are then used to refine these parameters by minimizing a cost function. The pipeline demonstrating the relationship between face regression and 3DMM is shown in Figure 1 Below, we briefly introduce the widely used FLAME model and the two latest FLAME-based regression tasks EMOCA and SMIRK.

FLAME [41] is a 3DMM used for synthesizing detailed and expressive 3D models of human heads. It accomplishes this by exploiting a linear shape space

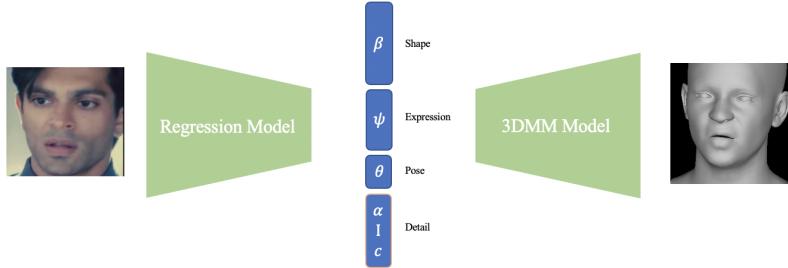


Fig. 1: A standard pipeline for 3D facial geometry reconstruction from an image. Left: The regression model extracts disentangled 3D parameter representations from the images. Right: These parameters are utilized to reconstruct the 3D facial geometry using a 3D Morphable Model.

trained from a large dataset of 3D scan sequences (4D with time). The model uses Principal Component Analysis (PCA) to create a low-dimensional representation of facial shapes and this involves identifying the principal components (PCs) that capture the most variance in the facial shapes from the training data. Thus, any new face can be represented as a linear combination (the shape parameters) of these principal components. By adjusting the weights of these components, FLAME can generate a wide variety of facial shapes. To improve its fidelity, FLAME also includes pose-dependent corrective blendshapes and global expression blendshapes. The global blendshapes are predefined facial expressions that are added to the base facial shape. They capture various facial movements associated with emotions, such as smiling, frowning, or surprise.

Figure 2 shows a few examples of the resulting synthesized faces when various model parameters are altered. The images were created from the FLAME Model Viewer located on the authors' site².

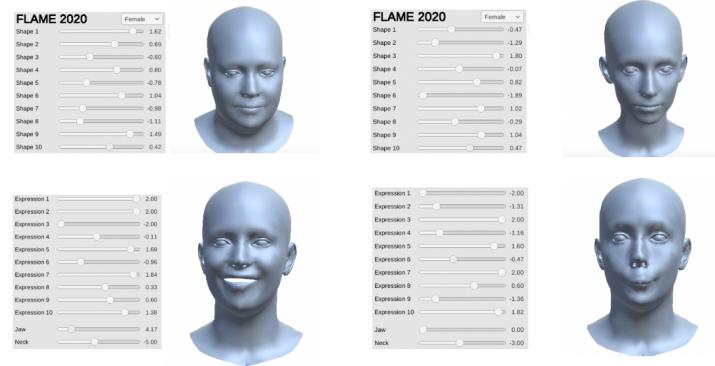


Fig. 2: 3D Representation Visualization

² <https://flame.is.tue.mpg.de/>

EMOCA [10] extends the previously trained face reconstruction framework, DECA [15], which reconstructs a detailed 3D face model by learning the FLAME parameters from a single RGB image. The model uses a combination of landmark losses, photometric losses, and perceptual emotion consistency loss during training to ensure the reconstructed 3D faces accurately convey the emotional content of the input images. The EMOCA model regresses a total of 334 parameters: 100 for shape, 50 for emotional expressions, 6 for pose, 100 for detail, 50 for texture, and others including pose-dependent and articulated components.

SMIRK [54] is another FLAME-based regression model that has the advantage of capturing any subtleties, extreme expressions, asymmetries, or rarely observed expressions that create slight deformations of the face shape. SMIRK replaces the previous differentiable rendering approach in comparing generated 3D face representations with the original inputs. Given the rendered predicted mesh geometry and sparsely sampled pixels of the input image, this new neural rendering module focuses on local geometries to generate a face image more similar to the original, which can then be fed back to the reconstruction pipeline. The SMIRK model regresses to 358 standard parameters of which 300 are shape, 50 are expression and 6 are pose. Other additional parameters include camera parameters and those specific to the neural rendering process used in SMIRK.

4 Fusion Architecture

4.1 3D Representation Classifier Architecture

Though EMOCA and SMIRK report emotion recognition performance, their results are based on a manually cleaned dataset, which is difficult to reproduce. Despite the inherent limitations of the AffectNet dataset, for a fair comparison, we extracted all 3D features from AffectNet datasets to curate a 3D representation dataset and then trained the 3D representation classifier.

The classifiers in this study adopted a similar architecture to the emotion classification MLP proposed in [60], with minor adjustments. The classifier network architecture comprises an input layer, followed by four fully connected layers, each with an output dimension of 2048. Batch normalization and Leaky ReLU activation functions are applied in all fully connected layers. Dropout rates of 50% and 40% are applied to the first and second layers, respectively. The output layer is adapted based on the dataset: for the RAF-DB dataset, it has 7 output dimensions, while for the AffectNet dataset, it includes 8 classes plus additional outputs for valence and arousal levels.

4.2 Loss Function

Discrete Expression Inference For the AffectNet dataset [47], following the previous approach [60, 62] for emotion classification, we employed a combined loss function for this task.

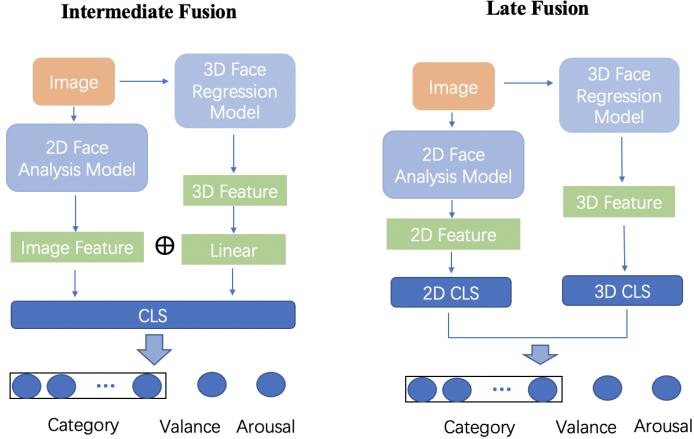


Fig. 3: Overview of the 3D Representation Fusion Architecture

The combined loss function used in this study integrates several components to handle both classification and regression tasks. The cross-entropy loss L_{CE} is used for the classification task, measuring the difference between the true and predicted class labels. For the regression tasks, the mean squared error L_{MSE} is applied to minimize the squared differences between the predicted and true values of valence and arousal. Additionally, we incorporate the Pearson correlation coefficient L_{PCC} , which assesses the linear correlation between the predicted and true valence and arousal values. Furthermore, the concordance correlation coefficient L_{CCC} is used to measure the agreement between the predicted and true values, considering both precision and accuracy.

The final loss function is a weighted combination of these components as shown Eq: 1:

$$Loss = L_{CE} + \frac{\alpha}{\alpha + \beta + \gamma} \times L_{MSE} + \frac{\beta}{\alpha + \beta + \gamma} \times (1 - L_{CCC}) + \frac{\gamma}{\alpha + \beta + \gamma} \times (1 - L_{PCC}) \quad (1)$$

where α, β, γ are weighting factors for the different components of the loss function. The values of α, β, γ are randomly sampled from a uniform distribution between 0 and 1 for each batch.

Continuous Expression Inference Regarding the VA estimation on AffectNet, we follow CAGE [62] and use a two-stage training approach and employ their best performance model, MaxViT as our 2D image analysis model. In the first stage, we train with a combined loss $Loss_{combined}$. In the second stage, building on the first, we modify the output to valence and arousal to strengthen the VA supervised learning with $Loss_{va}$.

$$Loss_{combined} = L_{weightedCE} + w_1 \cdot L_{MSE} \quad (2)$$

$$Loss_{va} = L_{CCC} + w_2 \cdot L_{MSE} \quad (3)$$

where the weighted cross-entropy $L_{weightedCE}$ is used to ease the issue of imbalanced distribution in the training set. This loss function assigns weights based on the frequencies of the expressions.

4.3 Intermediate Fusion Architecture

In the intermediate fusion architecture shown in Fig. 3, A facial image is processed by both a 2D face analysis model and a 3D face regression model. The 2D analysis model extracts high-dimensional features useful for recognizing facial expressions, while the 3D regression model regresses parameters that are linked to the facial structure.

Next, the 2D image features and the transformed 3D features are concatenated (or combined) into a comprehensive feature set. This feature set is then fed into a linear fusion architecture, leveraging both 2D and 3D information. The fused features are input into a classification layer (CLS), which is identical to the classifier architecture mentioned earlier. The CLS layer processes these features to predict various aspects of facial expressions (category, valence, or arousal).

Different datasets necessitate different outputs and loss functions. The loss function and linear fusion architecture used here are consistent with the classifier architecture mentioned above.

4.4 Late Fusion Architecture

In the late fusion architecture shown in Fig. 3, the input image is processed by the 2D face analysis model. This model extracts 2D features from the image and inputs these 2D features into the 2D classifier, which outputs emotion category, valence, and arousal results based on the requirements of different datasets. Meanwhile, the input image is also processed by the 3D face regression model. This model extracts 3D features from the image, and these 3D features are then input into a pre-trained classifier, which outputs emotion category, valence, and arousal results. This architecture maximizes the preservation of their respective independent inference capabilities.

In the late fusion step, the outputs of the 2D classifier and the 3D classifier are combined. Here, we use simple fusion methods such as max, mean, and weighted to integrate the information from both 2D and 3D analyses, generating more accurate and robust emotion recognition results.

5 Experiments

Training Setting and Evaluation Metrics The hyperparameters for the training architecture are summarized in Tab. 1. These settings ensured the stability and efficiency of the model training process.

Table 1: Hyperparameters for Classifier Training.

Hyperparameter	Value
Batch Size	64
Weight Decay	1e-5
Maximum Epochs	100
Early Stopping Patience	3
Learning Rate Scheduler	CyclicLR
Base Learning Rate	1e-6
Maximum Learning Rate	1e-4
Step Size	<code>len(train_loader)//2</code>
Scheduler Mode	'triangular'
Cyclical Momentum	No

For the emotion discrete classification task, we use state-of-the-art binary classification metrics [62, 71], including Accuracy, F1 score, Precision, and Recall. We follow established practices³ to compute these metrics. For a comprehensive evaluation, we report both weighted and macro averages for the unbalanced test dataset (RAF-DB). For the balanced test dataset (AffectNet), where both metrics are identical, we report results once.

For the emotion continuous regression task, our evaluation metrics are Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Concordance Correlation Coefficient (CCC).

Table 2: Classification Comparison of EMOCA and SMRIK 3D Representations only (no fusion) on AffectNet Dataset.

3D Classifier	Accuracy \uparrow	F1 \uparrow	Precision \uparrow	Recall \uparrow
$CLS_{Smirk3D-short}$	0.5461	0.5459	0.5477	0.5461
$CLS_{Smirk3D-full}$	0.5546	0.5547	0.5569	0.5546
$CLS_{Emoca3D-short}$	0.5723	0.5726	0.5758	0.5723
$CLS_{Emoca3D-full}$	0.5703	0.5704	0.5768	0.5703

5.1 3D Representation Classification Performance

In 3D representations, the parameters most relevant to expressions are expression, pose, and shape. To analyze the information gained from other parameters on FEI tasks, we divided them into two groups: the short group (expression, pose, shape) and the full group (all parameters). Thus, $Smirk3D_{short}$ has 353 dimensions, $Smirk3D_{full}$ has 358 dimensions, $Emoca3D_{short}$ has 156 dimensions, and $Emoca3D_{full}$ has 334 dimensions.

As shown in Tab. 2 and Tab. 3, $Emoca3D$ outperforms $Smirk3D$ in discrete emotion classification tasks. $Emoca3D_{short}$ achieves the best performance on AffectNet, improving by 2.26% in accuracy compared to $Smirk3D_{short}$, while $Smirk3D_{full}$ achieves the best performance on RAF-DB, improving by 3.7% in

³ Code reference for evaluation (see lines 90-92).

Table 3: Classification Comparison of EMOCA and SMRIK 3D Representations only (no fusion) on RAF-DB Dataset. Due to the unbalanced test dataset, we report both weighted and macro average metrics for a comprehensive evaluation. **Acc** stands for Accuracy, **F1** for F1 score, **P** for Precision, and **R** for Recall.

3D Classifier	Acc \uparrow	Weighted Avg			Macro Avg		
		F1 \uparrow	P \uparrow	R \uparrow	F1 \uparrow	P \uparrow	R \uparrow
$CLS_{Smirk3D-short}$	0.7378	0.7418	0.7475	0.7418	0.6421	0.6386	0.6482
$CLS_{Smirk3D-full}$	0.7557	0.7584	0.7631	0.7557	0.6585	0.6568	0.6627
$CLS_{Emoca3D-short}$	0.7862	0.7873	0.7895	0.7862	0.6965	0.6908	0.7037
$CLS_{Emoca3D-full}$	0.7927	0.7946	0.7985	0.7927	0.7073	0.7043	0.7118

accuracy compared to $Smirk3D_{full}$. Experiments show that EMOCA preserves more emotional information and is better suited for emotion reasoning tasks.

5.2 3D Fusion in Discrete Facial Expression Inference

In the experiments of discrete expression inference on the AffectNet dataset as shown in Tab. 4, our fusion method achieved the best performance. Although we did not replicate its best performance as reported (65.04%) on AffectNet 8, the weighted late fusion still achieved an improvement in all metrics. Here, the late fusion weight is 0.2 for the 3D component. The weighted fusion strategy improved the accuracy by 0.55%, the F1 score by 0.58%, the precision by 0.26%, and the recall by 0.55%.

Table 4: Classification Comparison of Different Fusion Architectures on AffectNet Dataset.

Framework	Accuracy \uparrow	F1 \uparrow	Precision \uparrow	Recall \uparrow
DDAMFN (our reproduction)	0.6324	0.6323	0.6353	0.6324
Intermediate Fusion				
$F_{2D} + F_{Smirk3D}$	0.6117	0.6098	0.6128	0.6117
$F_{2D} + F_{Emoca3D}$	0.6234	0.6232	0.6276	0.6234
Late Fusion				
Max with $CLS_{Smirk3D}$	0.6267	0.6260	0.6273	0.6267
Max with $CLS_{Emoca3D}$	0.6294	0.6292	0.6306	0.6294
Mean with $CLS_{Smirk3D}$	0.6262	0.6266	0.6315	0.6262
Mean with $CLS_{Emoca3D}$	0.6289	0.6295	0.6338	0.6289
Weighted with $CLS_{Smirk3D}$	0.6364	0.6367	0.6408	0.6364
Weighted with $CLS_{Emoca3D}$	0.6379	0.6381	0.6379	0.6379

The best performance was achieved with the weighted Emoca3D late fusion strategy, reaching the highest accuracy (94.00%), F1 score (93.93%), precision (93.97%), and recall (94.00%). Compared to our reproduced DDAMFN model, the weighted fusion strategy improved the accuracy by 3.84%, the F1 score by

Table 5: Classification Comparison of Different Fusion Architectures on RAF-DB Dataset. Due to the unbalanced test dataset, we report both weighted and macro average metrics for a comprehensive evaluation. **Acc** stands for Accuracy, **F1** for F1 score, **P** for Precision, and **R** for Recall.

Framework	Acc \uparrow	Weighted Avg			Macro Avg		
		F1 \uparrow	P \uparrow	R \uparrow	F1 \uparrow	P \uparrow	R \uparrow
DDAMFN (our reproduction)	0.9016	0.9013	0.9022	0.9016	0.8554	0.8686	0.8451
Intermediate Fusion							
$F_{2D} + F_{Smirk3D}$	0.9006	0.9007	0.9018	0.9006	0.8489	0.8561	0.8435
$F_{2D} + F_{Emoca3D}$	0.8996	0.8990	0.8989	0.8996	0.8501	0.8559	0.8453
Late Fusion							
Max with $CLS_{Smirk3D}$	0.8989	0.8984	0.8989	0.8989	0.8527	0.8656	0.8426
Max with $CLS_{Emoca3D}$	0.8941	0.8944	0.9021	0.8941	0.8462	0.8643	0.8485
Mean with $CLS_{Smirk3D}$	0.9030	0.9024	0.9041	0.9030	0.8561	0.8829	0.8361
Mean with $CLS_{Emoca3D}$	0.9130	0.9135	0.9178	0.9130	0.8413	0.8414	0.8521
Weighted with $CLS_{Smirk3D}$	0.9106	0.9099	0.9110	0.9106	0.8689	0.8914	0.8516
Weighted with $CLS_{Emoca3D}$	0.9400	0.9393	0.9397	0.9400	0.8958	0.9090	0.8860

3.80% the precision by 3.75%, and the recall by 3.84%. This result has made our model overpass the state-of-the-art performance on the RAF-DB dataset.

We believe the reason late fusion outperforms intermediate fusion is that the results from different models remain more independent, thereby maximizing the retention of each model’s respective advantages. However, the features provided by intermediate fusion are redundant compared to the original features. During training, we observed early stopping occurring within the first 8 epochs, earlier than the baseline model’s 12-15 epochs. We guess this is because intermediate fusion introduces excessive redundant information, which strongly correlates with features extracted from images in high-dimensional space, leading to reduced model generalization.

Table 6: Comparison with Previous SOTA models for Discrete FEI on RAF-DB Dataset.

Method	Accuracy [%]	Date [mm-yy]
FMAE [50]	93.09	07-2024
S2D [8]	92.57	12-2023
BTN [18]	92.54	07-2024
ARBEx [64]	92.37	05-2023
DDAMFN [71]	92.34	07-2023
Ours	94.00	07-2024

The results shown in Tab.6 indicate that a simple weighted late fusion strategy can significantly enhance the performance of our reproduced DDAMFN model (originally ranked 5th), achieving new state-of-the-art performance. Note,

that the TOP1 FMAE model pre-trained on the large dataset, while our method just trained with a single AffectNet dataset.

Table 7: Continuous VA Results from Different Fusion Architectures on AffectNet Dataset.

Framework	MSE ↓	MAE ↓	RMSE ↓	CCC ↑
<i>CAGE_{va}</i> (Our reproduction)	0.1044	0.2377	0.3230	0.7814
3D Representation				
<i>Regressor_{Emoca3D}</i>	0.1061	0.2483	0.3257	0.7637
Feature Fusion				
$F_{2D} + F_{Emoca3D}$	0.1061	0.2398	0.3257	0.7749
Late Fusion				
Max with <i>Regressor_{Emoca3D}</i>	0.1052	0.2419	0.3243	0.7727
Min with <i>Regressor_{Emoca3D}</i>	0.1053	0.2441	0.3245	0.7726
Mean with <i>Regressor_{Emoca3D}</i>	0.0956	0.2325	0.3092	0.7891
Weighted with <i>Regressor_{Emoca3D}</i>	0.0958	0.2316	0.3095	0.7901

Table 8: Benchmark Comparison for VA Inference on AffectNet Dataset.

Framework	RMSE _{val} ↓	RMSE _{aro} ↓	CCC _{val} ↑	CCC _{aro} ↑	Date[mm-yy]
VGG-G [5]	0.356	0.326	0.710	0.629	03-2021
CAGE [62]	0.331	0.305	0.716	0.642	04-2024
Ours	0.323	0.294	0.724	0.650	07-2024

5.3 3D Fusion in Continuous Facial Expression Inference

The experimental results on AffectNet, as shown in Table 7, indicate that our late fusion and mean fusion strategies effectively improved performance. Surprisingly, the performance of the 3D representation alone is already very close to that of our reproduced CAGE. Our mean fusion achieved an MSE of 0.0956 and an RMSE of 0.392, while our weighted fusion achieved an MAE of 0.2316 and a CCC of 0.7901. The weight for the 3D representation was set at 0.4. These results highlight the importance of 3D representation to continuous FEI tasks. Compared to our reproduced CAGE model, our mean late fusion improved performance by 8.43% in MSE, 2.19% in MAE, 4.27% in RMSE, and 0.99% in CCC. Our weighted late fusion increased performance by 8.24% in MSE, 2.57% in MAE, 4.18% in RMSE, and 1.11% in CCC.

Tab.8 shows that our late fusion method has surpassed the current state-of-the-art methods. Compared with the CAGE model, our valence RMSE increased by 2.42%, arousal RMSE increased by 3.61%, valence CCC increased by 1.12%,

and arousal CCV increased by 1.25%. The overall results indicate that the late fusion of 3D representations can effectively enhance continuous expression inference tasks.

The experimental results on the AffectNet as shown in Tab. 7, our late fusion and mean fusion both greatly improved the performance. Here the weight for 3D representation is 0.4. Our mean fusion achieved 0.0956 in MSE and 0.392 in RMSE, meanwhile, our weighted fusion achieved 0.2316 in MAE and 0.7901 in CCC. This highlights the 3D representation’s contribution to continuous FEI tasks. Our mean fusion achieved an MSE of 0.0956 and an RMSE of 0.392, while our weighted fusion achieved an MAE of 0.2316 and a CCC of 0.7901. Compared to our reproduced CAGE model, our mean late fusion performance improved by 8.43% in MSE, 2.19% in MAE, 4.27% in RMSE, and 0.99% in CCC. Our weighted late fusion performance increased by 8.24% in MSE, 2.57% in MAE, 4.18% in RMSE, and 1.11% in CCC.

6 Conclusions

As 3D face reconstruction aligns more closely with reality, its integration and analysis can be beneficial in many FEI tasks. Our research evaluates the performance of existing 3D face representations, introduces two fusion architectures, and demonstrates the efficiency of late fusion through extensive experiments. The experimental results show that our proposed method outperforms the state-of-the-art in AffectNet VA estimation and RAF-DB classification tasks. These findings offer valuable insights into the application of 3D representations for emotion inference.

Moving forward, our future work will focus on several key areas. First, we will delve deeper into the analysis of 3D representations within the realm of micro-expressions, aiming to capture even the most subtle emotional cues. Second, we will investigate how 3D reconstructions can enhance emotion inference in scenarios involving human backgrounds and human-human interactions. By continuing to refine and expand upon these methodologies, we aim to contribute to the development of more nuanced and accurate emotion inference models.

Acknowledgments

This work is based upon work supported under the National AI Research Institutes program by the National Science Foundation (NSF) and the Institute of Education Sciences (IES), U.S. Department of Education, through Award # 2229873 and NSF Award # 2223507. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF, the IES, or the U.S. Department of Education.

References

1. Antoniadis, P., Pikoulis, I., Filntysis, P.P., Maragos, P.: An audiovisual and contextual approach for categorical and continuous emotion recognition in-the-wild. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3645–3651 (2021)
2. Barrett, L.F., Adolphs, R., Marsella, S., Martinez, A.M., Pollak, S.D.: Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological science in the public interest* **20**(1), 1–68 (2019)
3. Benitez-Quiroz, C.F., Srinivasan, R., Martinez, A.M.: Emotionet: A large-scale, fine-grained dataset for automatic facial expression analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5562–5570 (2020). <https://doi.org/10.1109/CVPR42600.2020.00567>, <https://ieeexplore.ieee.org/document/9156975>
4. Bian, C., Zhang, Y., Yang, F., Bi, W., Lu, W.: Spontaneous facial expression database for academic emotion inference in online learning. *IET Computer Vision* **13**(3), 329–337 (2019)
5. Bulat, A., Cheng, S., Yang, J., Garbett, A., Sanchez, E., Tzimiropoulos, G.: Pre-training strategies and datasets for facial representation learning. In: European Conference on Computer Vision. pp. 107–125. Springer (2022)
6. Calvo, M.G., Lundqvist, D.: Facial expressions of emotion (kdef): Identification under different display-duration conditions. *Behavior research methods* **40**(1), 109–115 (2008)
7. Chen, K., Yang, X., Fan, C., Zhang, W., Ding, Y.: Semantic-rich facial emotional expression recognition. *IEEE Transactions on Affective Computing* **13**(4), 1906–1916 (2022)
8. Chen, Y., Li, J., Shan, S., Wang, M., Hong, R.: From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos. arXiv preprint arXiv:2312.05447 (2023)
9. Chen, Z., Feng, X., Zhang, S.: Emotion detection and face recognition of drivers in autonomous vehicles in iot platform. *Image and Vision Computing* **128**, 104569 (2022)
10. Danček, R., Black, M.J., Bolkart, T.: Emoca: Emotion driven monocular face capture and animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20311–20322 (2022)
11. Ding, C., Tao, D.: Robust face recognition via multimodal deep face representation. *IEEE transactions on Multimedia* **17**(11), 2049–2058 (2015)
12. Dong, L., Chaudhary, L., Xu, F., Wang, X., Lary, M., Nwogu, I.: Signavatar: Sign language 3d motion reconstruction and generation. arXiv preprint arXiv:2405.07974 (2024)
13. Ekman, P.: An argument for basic emotions. *Cognition & emotion* **6**(3-4), 169–200 (1992)
14. Epstein, S.: Controversial issues in emotion theory. *Review of Personality & Social Psychology* (1984)
15. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)* **40**(4), 1–13 (2021)
16. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.H., et al.: Challenges in representation learning: A report on three machine learning contests. *Neural Networks* **64**, 59–63 (2015)

17. Hall, D.L., Llinas, J.: An introduction to multisensor data fusion. *Proceedings of the IEEE* **85**(1), 6–23 (1997)
18. Her, M.B., Jeong, J., Song, H., Han, J.H.: Batch transformer: Look for attention in batch. *arXiv preprint arXiv:2407.04218* (2024)
19. Izard, C.E.: Basic emotions, relations among emotions, and emotion-cognition relations. (1992)
20. Kahou, S.E., Pal, C., Bouthillier, X., Froumenty, P., Gülcühre, Ç., Memisevic, R., Vincent, P., Courville, A., Bengio, Y., Ferrari, R.C., et al.: Combining modality specific deep neural networks for emotion recognition in video. In: *Proceedings of the 15th ACM on International conference on multimodal interaction*. pp. 543–550 (2013)
21. Khan, R.A., Crenn, A., Meyer, A., Bouakaz, S.: A novel database of children’s spontaneous facial expressions (liris-cse). *Image and Vision Computing* **83**, 61–69 (2019)
22. Kollias, D., Schulc, A., Hajiyev, E., Zafeiriou, S.: Analysing affective behavior in the first abaw 2020 competition. In: *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*. pp. 794–800 (2020)
23. Kollias, D.: Abaw: Valence-arousal estimation, expression recognition, action unit detection & multi-task learning challenges. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 2328–2336 (2022)
24. Kollias, D.: Abaw: learning from synthetic data & multi-task learning challenges. In: *European Conference on Computer Vision*. pp. 157–172. Springer (2023)
25. Kollias, D.: Multi-label compound expression recognition: C-expr database & network. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5589–5598 (2023)
26. Kollias, D., Sharmancka, V., Zafeiriou, S.: Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111* (2019)
27. Kollias, D., Sharmancka, V., Zafeiriou, S.: Distribution matching for heterogeneous multi-task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790* (2021)
28. Kollias, D., Tzirakis, P., Baird, A., Cowen, A., Zafeiriou, S.: Abaw: Valence-arousal estimation, expression recognition, action unit detection & emotional reaction intensity estimation challenges. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5888–5897 (2023)
29. Kollias, D., Tzirakis, P., Cowen, A., Zafeiriou, S., Kotsia, I., Baird, A., Gagne, C., Shao, C., Hu, G.: The 6th affective behavior analysis in-the-wild (abaw) competition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4587–4598 (2024)
30. Kollias, D., Tzirakis, P., Nicolaou, M.A., Papaioannou, A., Zhao, G., Schuller, B., Kotsia, I., Zafeiriou, S.: Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision* pp. 1–23 (2019)
31. Kollias, D., Zafeiriou, S.: Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface. *arXiv preprint arXiv:1910.04855* (2019)
32. Kollias, D., Zafeiriou, S.: Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792* (2021)
33. Kollias, D., Zafeiriou, S.: Analysing affective behavior in the second abaw2 competition. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 3652–3660 (2021)

34. Kollias, D., Zafeiriou, S., Kotsia, I., Dhall, A., Ghosh, S., Shao, C., Hu, G.: 7th abaw competition: Multi-task learning and compound expression recognition. arXiv preprint arXiv:2407.03835 (2024)
35. Kosti, R., Alvarez, J.M., Recasens, A., Lapedriza, A.: Emotion recognition in context based on emotic dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 889–895. IEEE (2017). <https://doi.org/10.1109/CVPRW.2017.112>, <https://arxiv.org/abs/2003.13401>
36. Kuncheva, L.I.: Combining pattern classifiers: methods and algorithms. John Wiley & Sons (2014)
37. Lahat, D., Adali, T., Jutten, C.: Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE* **103**(9), 1449–1477 (2015)
38. Li, S., Deng, W.: Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. *IEEE Transactions on Image Processing* **28**(1), 356–370 (2019)
39. Li, S., Deng, W., Du, J.: Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2584–2593. IEEE (2017)
40. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.* **36**(6), 194–1 (2017)
41. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* **36**(6), 194:1–194:17 (2017), <https://doi.org/10.1145/3130800.3130813>
42. Li, Z., Zhang, T., Jing, X., Wang, Y.: Facial expression-based analysis on emotion correlations, hotspots, and potential occurrence of urban crimes. *Alexandria Engineering Journal* **60**(1), 1411–1420 (2021)
43. Liu, Y., Dai, W., Feng, C., Wang, W., Yin, G., Zeng, J., Shan, S.: Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3688–3697 (2022). <https://doi.org/10.1109/ICCVW54120.2022.00108>, <https://arxiv.org/abs/2208.00847>
44. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: 2010 ieee computer society conference on computer vision and pattern recognition-workshops. pp. 94–101. IEEE (2010)
45. Lyons, M.J., Kamachi, M., Gyoba, J.: The japanese female facial expression (jaffe) database. Zenodo (1998). <https://doi.org/10.5281/zenodo.14140>, <https://zenodo.org/record/14140>
46. Mehrabian, A., Wiener, M.: Decoding of inconsistent communications. *Journal of personality and social psychology* **6**(1), 109 (1967)
47. Mollahosseini, A., Hasani, B., Mahoor, M.H.: Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* (2017)
48. Munsif, M., Sajjad, M., Ullah, M., Tarekegn, A.N., Cheikh, F.A., Tsakanikas, P., Muhammad, K.: Optimized efficient attention-based network for facial expressions analysis in neurological health care. *Computers in Biology and Medicine* **179**, 108822 (2024)
49. Nan, Y., Ju, J., Hua, Q., Zhang, H., Wang, B.: A-mobilenet: An approach of facial expression recognition. *Alexandria Engineering Journal* **61**(6), 4435–4444 (2022)

50. Ning, M., Salah, A.A., Ertugrul, I.O.: Representation learning and identity adversarial training for facial behavior understanding. arXiv preprint arXiv:2407.11243 (2024)
51. Paper with code. <https://paperswithcode.com/dataset/affectnet> (2024), accessed: 2024-08-07
52. Prentice, C., Dominique Lopes, S., Wang, X.: Emotional intelligence or artificial intelligence—an employee perspective. *Journal of Hospitality Marketing & Management* **29**(4), 377–403 (2020)
53. Ren, Z., Ortega, J., Wang, Y., Chen, Z., Guo, Y., Yu, S.X., Whitney, D.: Veatic: Video-based emotion and affect tracking in context dataset. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 4467–4477 (2024)
54. Retsinas, G., Filntisis, P.P., Danecek, R., Abrevaya, V.F., Roussos, A., Bolkart, T., Maragos, P.: 3d facial expressions through analysis-by-neural-synthesis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 2490–2501 (2024)
55. Roseman, I.J.: Cognitive determinants of emotion: A structural theory. *Review of personality & social psychology* (1984)
56. Russell, J.A.: A circumplex model of affect. *Journal of personality and social psychology* **39**(6), 1161 (1980)
57. Shan, M., Dong, L., Han, Y., Yao, Y., Liu, T., Nwogu, I., Qi, G.J., Hill, M.: Towards open domain text-driven synthesis of multi-person motions. arXiv preprint arXiv:2405.18483 (2024)
58. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition. In: *Proceedings of the Neural Information Processing Systems (NIPS)*. vol. 11 (2015)
59. Tan, M., Le, Q.: Efficientnetv2: Smaller models and faster training. In: *International conference on machine learning*. pp. 10096–10106. PMLR (2021)
60. Toisoul, A., Kossaifi, J., Bulat, A., Tzimiropoulos, G., Pantic, M.: Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence* **3**(1), 42–50 (2021)
61. Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y.: Maxvit: Multi-axis vision transformer. In: *European conference on computer vision*. pp. 459–479. Springer (2022)
62. Wagner, N., Mätzler, F., Vossberg, S.R., Schneider, H., Pavlitska, S., Zöllner, J.M.: Cage: Circumplex affect guided expression inference. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4683–4692 (2024)
63. Wang, J., Li, Z., Liu, Y., Fu, Q.: Semantic-rich facial emotional expression recognition with emo135 dataset. In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. pp. 1150–1157 (2021). <https://doi.org/10.1609/aaai.v35i2.16214>, <https://ojs.aaai.org/index.php/AAAI/article/view/16214>
64. Wasi, A.T., Šerbetar, K., Islam, R., Rafi, T.H., Chae, D.K.: Arbex: Attentive feature extraction with reliability balancing for robust facial expression learning. arXiv preprint arXiv:2305.01486 (2023)
65. Wu, D., Pigou, L., Kindermans, P.J., Le, N.D.H., Shao, L., Dambre, J., Odobez, J.M.: Deep dynamic neural networks for multimodal gesture segmentation and recognition. *IEEE transactions on pattern analysis and machine intelligence* **38**(8), 1583–1597 (2016)

66. Xu, F., Chaudhary, L., Dong, L., Setlur, S., Govindaraju, V., Nwogu, I.: A comparative study of video-based human representations for american sign language alphabet generation. In: 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–6. IEEE (2024)
67. Yi, D., Lei, Z., Li, S.Z.: Shared representation learning for heterogenous face recognition. In: 2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG). vol. 1, pp. 1–7. IEEE (2015)
68. Yun, S.S., Choi, J., Park, S.K., Bong, G.Y., Yoo, H.: Social skills training for children with autism spectrum disorder using a robotic behavioral intervention system. *Autism Research* **10**(7), 1306–1323 (2017)
69. Zafeiriou, S., Kollias, D., Nicolaou, M.A., Papaioannou, A., Zhao, G., Kotsia, I.: Aff-wild: Valence and arousal ‘in-the-wild’challenge. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on. pp. 1980–1987. IEEE (2017)
70. Zhai, Y., Huang, M., Luan, T., Dong, L., Nwogu, I., Lyu, S., Doermann, D., Yuan, J.: Language-guided human motion synthesis with atomic actions. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 5262–5271 (2023)
71. Zhang, S., Zhang, Y., Zhang, Y., Wang, Y., Song, Z.: A dual-direction attention mixed feature network for facial expression recognition. *Electronics* **12**(17), 3595 (2023)
72. Zhang, W., Liu, Y., Chen, J., Qiu, X.: Fine-grained facial expression database (f2ed) for emotion perception and expression recognition. In: Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW). pp. 1720–1728 (2020). <https://doi.org/10.1109/ICCVW50498.2020.00204>, <https://ieeexplore.ieee.org/document/9200887>