3D Face Reconstruction with Graph Attention Based on Action Unit Features

Hyeonjin Kim
Dept. of Computer Engineering
Kwangwoon University
Seoul, Korea
zx8635@gmail.com

Hyukjoon Lee
Dept. of Computer Engineering
Kwangwoon University
Seoul, Korea
hlee@kw.ac.kr

Jinheung Kong
Dept. of Embedded Software Enngeeing
Kwangwoon University
Seoul, Korea
kongjh@kw.ac.kr

***Abstract*—The reconstruction of 3D face shapes and expressions from 2D images remains unconquered due to the lack of detailed modeling of human facial movement based on the correlation between the different parts of faces. We propose to use facial action units (AUs), which are detailed taxonomy of the human facial movements based on observation of activation of muscles or muscle groups, in the 3D facial expression reconstruction to allow for detailed modeling of various facial expression types. We present a novel 3D face reconstruction framework called Action Unit feature-based Graph Attention Network Encoder (AUGANE) that can generate a 3D face model that is responsive to AU activation given a single monocular 2D image to capture expressions. AUGANE leverages AU-specific features as well as facial global features to enable accurate 3D facial expression reconstruction with Graph Attention Network Encoder (GANE). We also introduce a novel loss function which is to force learning toward the minimal discrepancy in AU activation between the input and rendered reconstruction. The experimental results demonstrate the superior performance of the proposed framework over state-of-the-art methods.**

Keywords—3D face reconstruction, action units, graph attention, transformer, 3DMM

I. INTRODUCTION

In recent years, rapid advances in deep learning technology have led to numerous innovative achievements in computer vision and graphics research. 3D face reconstruction from 2D images has received a tremendous amount of attention in computer vision and has made major progresses thanks to the highly accurate modeling capability of deep neural networks. 3D face reconstruction enables various applications such as speech-driven 3D facial animation, 3D avatar generation, virtual makeup, performance capture, virtual and augmented reality, and human-robot interaction [1][2][3][4][5][6].

Most existing studies use pre-computed 3D morphable models (3DMMs) to incorporate prior knowledge about facial geometry and appearance [7]. These methods take advantage of the rich information stored in 3DMM to improve the accuracy and fidelity of reconstructed 3D faces [8]. In recent studies, deep neural networks have been employed to predict the parameter values of 3DMM based on self-supervised learning, which project the reconstructed 3D face into the image plane to generate rendered image and calculate various loss functions such as landmark reprojection loss, face recognition loss, and photometric loss [9][10][11].

In recent studies, it has been pointed out that depending on such loss functions only is insufficient for capturing rich and subtle facial expressions [12]. EMOCA employed an emotion consistency loss function which computes the distance between the emotion recognition network outputs of input and rendered image during the training process in order to ensure that the two images convey emotions that are perceptually similar. SPECTER employed a perceptual lip movement loss function that can express visual speech information as a 3D face by applying lip reading recognition to input and rendered images [13].

Facial AUs are detailed taxonomy of the human facial movements and defined based on observation of activation of a muscle or muscle group. Unlike categorized emotion models, AUs provide comprehensive and objective means to characterize human facial expressions [14][15]. Thus, we try to include AUs in the 3D facial expression reconstruction for the detailed modeling of various facial expression types.

In this paper, we propose a novel 3D face reconstruction framework called AU feature-based Graph Attention Network Encoder (AUGANE). AUGANE utilizes features generated from a pretrained AU-specific Feature Generator (AFG) of state-of-the-art AU detection framework, named ME-GraphAU [15], and the global facial features generated from a pretrained 3D face reconstruction network, named DECA [9]. We introduce the Graph Attention Network Encoder (GANE)-based 3D face reconstruction model to predict 3D face reconstruction parameter values from these features. GANE learns the relationships between these generated features through graph attention mechanism with Graph Attention Network (GAT), and predict 3D face reconstruction parameter values with transformer encoder. In addition, we introduce novel AU-based loss functions such as AU-weighted landmark reprojection loss function, AU-based relative distance loss function, and AU feature loss function which is to force learning toward the minimal discrepancy in AU activations between the face in an input image and the reconstructed 3D face.

This paper is organized as follows: In section 2, a brief introduction to some background knowledge on 3D face models, 3D face reconstruction and facial AUs are provided. In section 3, the proposed framework is explained in detail followed by the experimental results. Finally, we end our discussion with the concluding remarks.

II. BACKGROUNDS

A. 3D Face Models

Vetter and Blantz explained a method for reconstructing a 3D face from a single image with a pre-computed 3DMM in an analysis-by-synthesis fashion [7]. 3DMM is statistical models capable of capturing and representing various facial changes in low-dimensional space. These models are built from a vast amount of 3D facial scan data. The traditional 3DMM was based on Principal Component Analysis (PCA) for facial shape, but recent models such as FLAME, Basel Face Model, FaceWarehouse have separated shape, expression, and appearance spaces, enabling richer representations [7][8].

FLAME was trained on 33,000 scan data and represents shape, pose, and expression parameters in well-separated spaces through an effective parameter separation process. FLAME consists of a template mesh, shape blendshapes, pose blendshapes, and expression blendshapes. Each blendshape is composed of displacements from the template mesh, with PCA applied to shape and expression. They applied an iterative optimization approach for each parameter during the model training process to separate the spaces of each parameter. Through the combination of parameter space separation and the utilization of multiple scan data, FLAME enables easier and more accurate facial reconstruction compared to other 3DMM models. For this reason, FLAME is by far the most widely used choice for tasks involving 3D faces. In this paper, we leverage FLAME as a powerful and expressive tool in modeling facial geometry and expressions.

B. 3D Face Reconstruction

In 3D face reconstruction research using 3DMM, it is common to estimate model parameters most suitable for RGB images. The direct optimization procedure is mainly carried out through an analysis-by-synthesis framework to estimate the model parameters. Optimization-based frameworks were used in early studies. However, optimization-based methods were later replaced by deep learning-based approaches due to the computational complexity and long processing times for each image. Deep learning-based methods that directly learn the mapping between 2D images and 3D faces have grown rapidly over the last few years, becoming a standard choice to replace a wide range of statistical model fitting [9][10].

Early deep learning-based 3D face reconstruction methods faced challenges related to the training dataset and training strategies. They were required to collect numerous 3D facial scan data corresponding to 2D images to train a deep learning-based model. However, this was hampered by the cost and inefficiency of obtaining numerous 3D facial scan data. A self-supervised learning framework was proposed that minimizes the difference between input images and rendered images to address this issue. The self-supervised learning framework utilizes a differentiable renderer to directly calculate the difference between input and rendered images, enabling end-to-end learning [17]. For effective optimization of the self-supervised learning approach in 3D face reconstruction, a training strategy is essential. RingNet [11] and DECA [Feng *et al*., 2021] applied a landmark-based training strategy by predicting landmarks for input images and using them indirectly as pseudo ground truth. They use landmark reprojection loss which computes the distance between the ground-truth 2D face landmark and its corresponding landmark on the surface of the 3DMM, projected onto the image. Additionally, EMOCA [12] employed a perception-based training strategy by utilizing a deep learning-based emotion recognition model as a feature extractor to minimize the distance of features for input and rendered images.

C. Facial AUs

AU detection involves analyzing facial expressions to detect independent movements in each region of the face [16]. Actual facial movements and expression styles vary widely across individuals [14]. The Facial Action Coding System (FACS) was developed to represent human expressions independently of each individual [18]. FACS encodes facial movements into AUs based on the observations of activation of the facial muscles or muscle groups.

Research on automated AU detection has been actively conducted, which is useful in tasks related to image-based facial behavior analysis [19]. AU detection can be formulated as a multi-label classification problem, and research based on machine learning and deep learning for this task has been actively conducted. In addition, each AU has underlying relationships emphasizing the need to consider these relationships in AU detection research [20]. ME-Graph AU [15] utilized a Convolutional Neural Network (CNN) and Graph Neural Network (GNN)-based model for AU detection, considering the relationships between AUs. Initially, a CNN-based network generates a facial representation for the input image. The AU-specific Feature Generator (AFG), composed of Fully Connected layers (FC layer) and Global Average Pooling layer (GAP layer), extracts AU-specific features from the overall facial representation. To model the relationships between the extracted AU features, a GNN-based network produces an AU relation graph. The AU relation graph includes relationships for each pair of AUs and predicts the activation probabilities and co-occurrence patterns of AUs. ME-GraphAU demonstrated state-of-the-art performance in AU detection benchmarks BP4D [21] and DISFA [22]. In this paper, we apply these AU characteristics to 3D face reconstruction, enhancing the performance of 3D expression representation.

III. AUGANE

A. Architecture

Figure 1 shows the overall architecture AUGANE framework. AUGANE learns the relation graph among AU-specific features and the global facial representation to predict accurate 3D face reconstruction parameter values. The activation of AUs has an individual relationship with each other and describes the overall facial expression [15][18]. We represent and model the relationships among the AU-specific features and the global facial features by using a graph structure, inspired by the previous research works in [20][23] which consider the AU features only.

We employ the pre-trained AFG block from ME-GraphAU to generate the AU-specific features from the face in an image. The AFG is encouraged to generate the AU-specific features dedicated to the AU detection model. The AU-specific features contain both AU activation status and their associations for each facial display. These features can provide a richer representation of subtle details in facial expressions. The AFG takes an input image and generates AU-specific features as:

$$\boldsymbol{V_{AFG}} = \{\vec{\boldsymbol{v}}_{\boldsymbol{1}}, \vec{\boldsymbol{v}}_{\boldsymbol{2}}, \dots, \vec{\boldsymbol{v}}_{\boldsymbol{N}}\}, \vec{\boldsymbol{v}}_{\boldsymbol{i}} \in \mathbb{R}^{\boldsymbol{512}}, \qquad (1)$$

where N is the number of AU-specific features. The DECA pretrained 3D face reconstruction model is used as a facial global feature generator. The DECA encoder is composed of a CNN and a FC layer. The CNN extracts the global face representation $\boldsymbol{X_{DECA}} \in \mathbb{R}^{\boldsymbol{2048}}$ while the FC layer generates the 3D face reconstruction parameters $\boldsymbol{\Theta_{DECA}} \in \mathbb{R}^{\boldsymbol{236}}$ from $\boldsymbol{X_{DECA}}$. The global face representation $\boldsymbol{X_{DECA}}$ contains generalized global features of a face in the input image. The AU-specific features V_{AFG} and the facial global features $\boldsymbol{X_{DECA}}$ are concatenated to form the input for GANE such that:

$$\boldsymbol{V_{GANE}} = \boldsymbol{V_{AFG}} \parallel \boldsymbol{X_{DECA}}\boldsymbol{E}, \boldsymbol{E} \in \mathbb{R}^{\boldsymbol{2048\times512}} \qquad (2)$$

where $\parallel$ is the concatenation operator, and $\boldsymbol{E}$ is the trainable linear projection. The overall procedure of

generating $\boldsymbol{V_{GANE}}$ from an input with AFG and DECA is illustrated in Figure 1.

We propose GANE, a deep learning model on graph-structured data, which regresses 3D face reconstruction parameter values from generated node features $\boldsymbol{V_{GANE}}$. GANE is composed of a GAT for learning relationships within nodes and a transformer encoder for predicting 3D face reconstruction parameter values from the captured relation graph and generated features. Since the underlying relationship present in the activation of AUs is independent from each other, it is appropriate to model their relationships through GAT. The GAT enables flexible representation of the graph by dynamically assigning weights and can capture the different importance between neighboring nodes [24]. This capability makes it suitable for handling the asymmetric characteristics of activation of AUs. The GAT takes $\boldsymbol{V_{GANE}}$ as its input and produces the attention embedded node feature such that:

$$\boldsymbol{V'_{GANE}} = \{\vec{v}'_1, \vec{v}'_2, \dots, \vec{v}'_N, \vec{v}'_{N+1}\}, \vec{v}'_i \in \mathbb{R}^{512}. \quad (3)$$

The generated features represent visual features related to specific or global parts of the face. Following the approach by a previous study, where they show that the transformer can operate on patches extracted from irregular grids, allowing for the utilization of visual tokens from irregular facial regions without the need for uniform spaced sampling [25], we input AU-specific features and global facial features into a transformer encoder to predict expression parameter values for 3D face reconstruction. The transformer encoder receives $\boldsymbol{V'_{GANE}}$ and then maps $\boldsymbol{V'_{GANE}}$ to 3D face reconstruction parameters $\boldsymbol{\Theta_{GANE}}$. The trainable parameter query token $\boldsymbol{T_Q}$ and $\boldsymbol{V'_{GANE}}$ are concatenated and fed into the input of the transformer encoder. Figure 1(c) shows how GANE generates 3D face reconstruction parameter values.

The 3D face reconstruction parameter values $\boldsymbol{\Theta_{GANE}}$ are handed over to the FLAME decoder for the 3D face reconstruction. The differentiable renderer then renders the final image from the reconstructed 3D face. The differentiable renderer enables end-to-end training. Finally, the loss between the input image $\boldsymbol{I}$ and the rendered image $\boldsymbol{I_{Re}}$ is calculated to train the proposed model.

B. Loss Functions

GANE is trained by minimizing the total loss:

$$\boldsymbol{L_{total}} = \boldsymbol{L_{auLmk}} + \boldsymbol{L_{auRel}} + \boldsymbol{L_{auFeat}} + \boldsymbol{L_{reg}}. \quad (4)$$

We explain below each of these loss functions in more detail.

*a) **AU-weighted landmark reprojection loss**:* The movement of landmarks triggered by the activation of AUs is known to serve as an effective means to describe the AUs [26]. $\boldsymbol{L_{auLmk}}$ assigns dynamic weights to the facial regions where AUs are activated to encourage the accurate representation of AUs in the reconstructed face. This enables GANE to pay more attention to activated AUs during the training process. The AU-weighted landmark reprojection loss function is defined as:

$$\boldsymbol{L_{auLmk}} = \sum_{i=1}^{N} \sum_{j=1}^{L_i} \boldsymbol{p_i} \left\| \boldsymbol{k_j} - \boldsymbol{s\Pi}(\boldsymbol{M_j}) + \boldsymbol{t} \right\|_1, \quad (5)$$

where $\boldsymbol{N}$ is the number of AUs, $\boldsymbol{L_i}$ is the number of landmarks related to $\boldsymbol{i^{th}}$ AU, $\boldsymbol{p_i}$ is the activation status of the $\boldsymbol{i^{th}}$ AU predicted by ME-GraphAU, $\boldsymbol{k_j}$ is the $\boldsymbol{j^{th}}$ landmark coordinate in the input image and the $\boldsymbol{M_j}$ is corresponding landmark on the FLAME model's surface. $\boldsymbol{s, \Pi, t}$ represent the predicted camera parameters, denoting the isotropic scale $\boldsymbol{s}$, orthographic 3D-2D projection matrix $\boldsymbol{\Pi}$, and 2D transition $\boldsymbol{t}$, respectively. The Mediapipe [27] landmark detector is used to predict landmarks from 2D images based on a total of 105 landmarks distributed across the eyebrows, eyes, nose, and mouth regions.

*b) **AU-based relative distance loss**:* The AU-based relative distance loss computes the relative distance between AU configural features for image landmarks and the

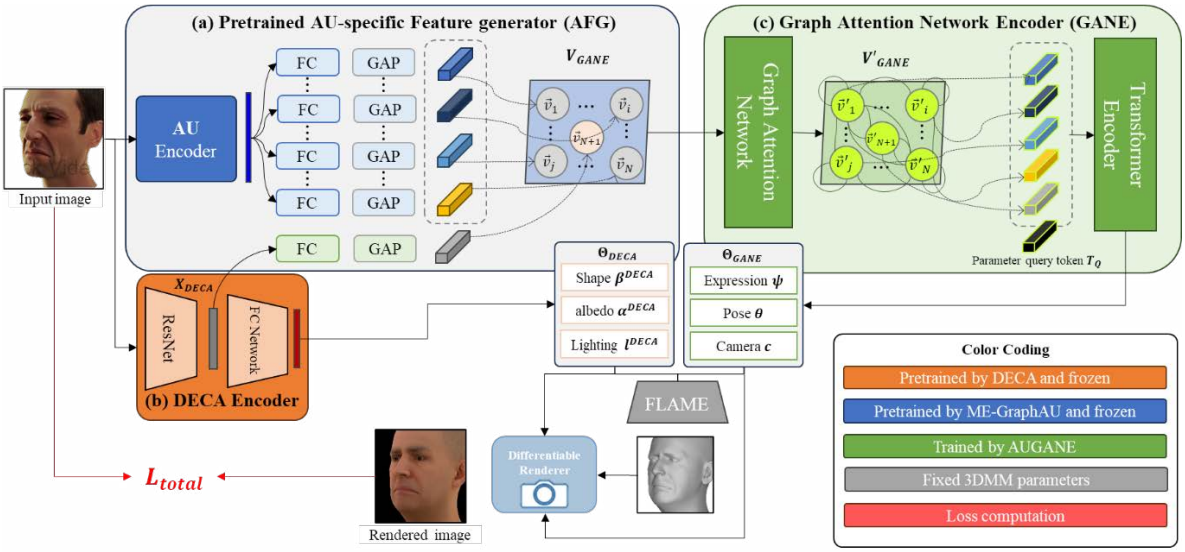


Figure 1 Overview of architecture for AU-guided 3D face reconstruction with AUGANE. (a) Pretrained AU-specific feature generator receives input image and generate AU-specific features. (b) Pretrained DECA encoder receives input image and generate global facial features as well as 3D face reconstruction parameter values. (c) Our GANE-based model receives both AU-specific features and global facial features and predict 3D face reconstruction parameter values.

projected 3D landmarks. The AU configural features involve calculating relative distances between facial landmark points and are used to determine AUs [28]. For example, Brow Lowerer AU is determined based on the distance between the landmark points 21 and 22, which correspond to the inner eyebrow landmarks on the left and right. This type of loss function is similar to eye closure loss of DECA, which computes an error in the relative offset between landmarks on the upper and lower eyelids for image landmarks and their corresponding projected 3D landmarks. We extend this approach in the context of AU by incorporating configural features. The AU-based relative distance loss computes the errors in configural features of image landmarks $\boldsymbol{k}$ and corresponding 3D landmarks $\boldsymbol{M}$ projected onto the image plane:

$$\boldsymbol{L_{auRel}} = \sum_{i=1}^{23} \left\| \boldsymbol{c_i^k} - \boldsymbol{c_i^{s\Pi(M)}} \right\|_1 \quad (6)$$

where $\boldsymbol{c_i^k}$ and $\boldsymbol{c_i^{s\Pi(M)}}$ are $\boldsymbol{i^{th}}$ configural features of image landmarks $\boldsymbol{k}$ and projected 3D landmarks $\boldsymbol{s\Pi(M)}$. The proposed configural features from [28] are defined using 66 landmarks model, but we modify landmark model with 105 landmarks from HRNet. The landmark indices and configural features corresponding to each AU are described in Table 1.

*c) **AU feature loss**:* The AU feature loss computes the distances between the AU-specific features of the input image $\boldsymbol{I}$ and the rendered image $\boldsymbol{I_{Re}}$:

$$\boldsymbol{L_{auFeat}} = \|\boldsymbol{AFG(I)} - \boldsymbol{AFG(I_{Re})}\|_2. \quad (7)$$

Optimizing this loss during training encourages the reconstructed 3D face to convey AU activations that are visually similar to the image.

*d) **Parameter regularizer.*** $\boldsymbol{L_{reg}}$ regularizes expression $\boldsymbol{\psi}$, pose $\boldsymbol{\theta}$, camera parameters $\mathbf{c}$ and is formulated as:

$$\boldsymbol{L_{reg}} = \|\boldsymbol{\psi}\|_2^2 + \|\boldsymbol{\theta}\|_2^2 + \|\boldsymbol{c}\|_2^2 \quad (8)$$

IV. Experiments

A. Implementation Details

AUGANE is trained with a total of approximately 300,000 images from VGGFace2, Aff-wild2, CelebA-HQ, FFHQ, and BUPT-CB [29][30][31][32][33]. We use PyTorch3D to render the reconstructed 3D face onto the image plane. In addition, we use the Adam optimizer to optimize parameters during the training process, and the learning rate is 1e-05, the batch size is 16, and the epoch is 15. For parameter regularization, 1e-05 is applied to the expression parameter, and 0.1 is applied to the pose parameter. The loss function weighting parameters for each loss function is 0.75 for the $\boldsymbol{L_{auLmk}}$, 0.25 for the $\boldsymbol{L_{auRel}}$, and 0.75 for the $\boldsymbol{L_{auFeat}}$. The GANE model predicts only the expression $\boldsymbol{\psi}$, pose $\boldsymbol{\theta}$, and camera $\boldsymbol{c}$ parameter values among the 3D face reconstruction parameters, and DECA predicts the shape $\boldsymbol{\beta^{DECA}}$, light $\boldsymbol{l^{DECA}}$, albedo $\boldsymbol{a^{DECA}}$ parameter values.

B. Quantatative Evaluations

While standard benchmarks exist for quantitative evaluating the identity face shape in the context of 3D face reconstruction [34], there is currently no benchmark specifically designed to evaluate the performance of the expression reconstruction methods.

TABLE 1. FACIAL AUS AND CORRESPONDING FACIAL LANDMARKS

Facial parts	Related AUs	Involved landmarks
Brow	Brow Lowerer	0, 1, 2, …, 19
Inner brow	Inner Brow Raiser	1, 3, 5, 6, 8, 9, 11, 13, 15, 16, 18, 19
Outer brow	Outer Brow Raiser	Elements excluding ***Inner brow*** from ***Brow***
Eye	Lid Tightener	20, 21, …, 51
Lower eye	Cheek Raiser	20, 21, …, 27, 33, 36, 37, …, 43, 49
Upper eye	Upper Lid Raiser	Elements excluding ***Lower eye*** from ***Eye***
Nose	Nose Wrinkler	52, 53, …, 64
Mouth	Lip Pucker, Lip Stretch, Lip Tightener	65, 66, …, 104
Upper mouth	Upper Lip Raiser	65, 66, 69, 70, …, 76, 85, 86, …, 94, 103, 104
Mouth corner	Lip Corner Puller, Lip Corner Depressor	71, 72, 73, 74, 79, 80, 81, 82, 85, 86, 88, 89, 90, 91, 92, 93, 97, 98, 99, 100, 103, 104

When evaluating the expression of a 3D face by measuring the difference between the reconstructed 3D face with a ground-truth scan, the error from the scan is significantly influenced by the facial identity. As a result, we objectively evaluate the methods through a comparison of AU activation states detected by ME-GraphAU between input images and rendered images. We employ the DISFA dataset [22] for quantitative evaluation, as it serves as one of the training datasets for ME-GraphAU. This choice is made with the confidence that the AU detection model will effectively detect AU activation states during the evaluation. The DISFA dataset contains 27 subjects watching a video and consists of 130,815 frames. We select and evaluate with the top 9 videos based on the number of detected AU activations to select videos with diverse facial expressions among the 27 videos. The evaluation results for each subject, and evaluation results for each AU are reported in Tables 2 and 3.

In the per-subject evaluation results presented in Table 2, AUGANE outperforms both DECA and EMOCA for all subjects. Table 3 provides an evaluation per-AU. Among the evaluated AUs, AU1 (Inner Brow Raiser) and AU2 (Outer Brow Raiser) are observed through the contraction of the frontalis muscle, resulting in central forehead wrinkles and eyebrow movements [34]. However, within the scope of this paper, considering detailed reconstruction was not pursued. Detecting these AUs from 3D faces reconstructed by AUGANE, DECA, and EMOCA becomes challenging. Consequently, a F1 score of 0 is recorded for all methods. On the other hand, AUGANE demonstrated superior performance compared to both DECA and EMOCA for AUs related to the upper face (AU1, 2, 4, 5, 7). Particularly, for AU 4 (Brow Lowerer), it exhibited performance that was 5 times higher than DECA and 1.5 times higher than EMOCA. In addition, for the lower face, AUGANE surpassed DECA significantly and showed performance either surpassing or comparable to EMOCA for some AUs. In conclusion, the average F1 scores were 0.39, 0.18, and 0.30 for AUGANE, DECA, and EMOCA, respectively, confirming that AUGANE achieved the highest performance.

TABLE 2. PER-SUBJECT F1 SCORE EVALUATION RESULTS FOR AU DETECTION ON INPUT IMAGES AND RENDERED IMAGES

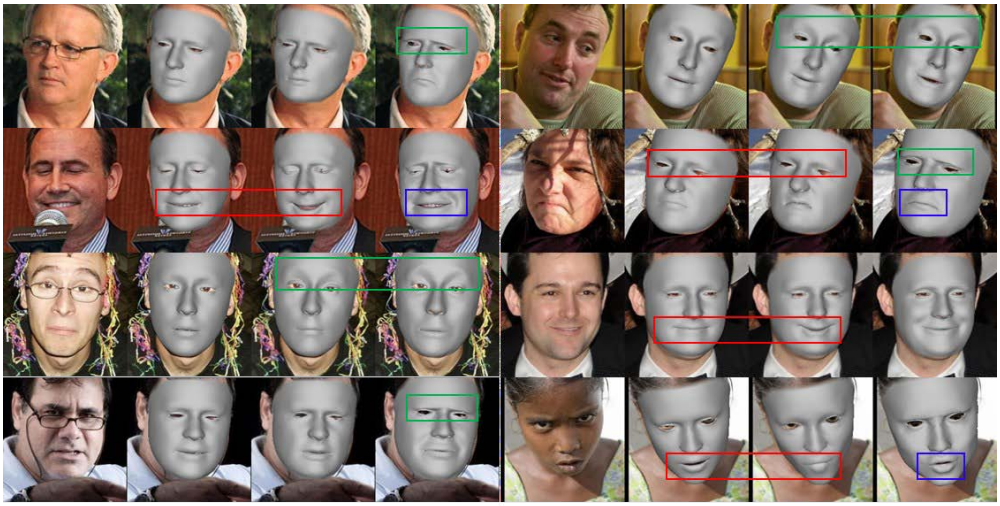
Subject	Method		
	AUGANE	DECA	EMOCA
03	**0.46**	0.23	0.24
06	**0.35**	0.13	0.32
11	**0.45**	0.18	0.30
12	**0.46**	0.30	0.43
16	**0.43**	0.05	0.24
18	**0.45**	0.19	0.25
23	**0.27**	0.16	0.24
25	**0.34**	0.21	0.25
27	**0.33**	0.07	0.30
Avg.	**0.39**	0.18	0.30

TABLE 3. PER-AU F1 SCORE EVALUATION RESULTS FOR AU DETECTION ON INPUT IMAGES AND RENDERED IMAGES

AU	Method		
	AUGANE	DECA	EMOCA
01	0.00	0.00	0.00
02	0.00	0.00	0.00
04	**0.47**	0.09	0.30
05	**0.24**	0.08	0.12
07	**0.78**	0.47	0.64
09	0.17	0.00	**0.21**
10	**0.83**	0.47	0.81
12	0.84	0.40	**0.86**
15	**0.13**	0.00	0.02
20	0.04	0.01	**0.06**
23	**0.28**	0.03	0.02
26	**0.56**	0.46	0.36
Avg.	**0.39**	0.18	0.30

C. *Qualatative Evaluations*

For qualitative evaluation, we employ 300W dataset and the DISFA dataset. Figure 2 shows the 3D face reconstruction results for the 300W dataset. We highlighted with red boxes when there were errors, and with green/blue boxes when the upper and lower faces were reconstructed accurately. Here,



(a) Input image (b) DECA (c) EMOCA (d) AUGANE (a) Input image (b) DECA (c) EMOCA (d) AUGANE

Figure 2 Visual comparison with DECA, EMOCA, and AUGANE on the 300W dataset.

AUGANE captures facial movements better than DECA and EMOCA. with more robust reconstruction performance for upper facial movements. In addition, AUGANE can express AU 12 (Lip Corner Puller) and AU 15 (Lip Corner Depressor) more accurately than DECA and EMOCA.

The experimental results for the DISFA dataset are presented in Figure 3. From left to right, two close frames are shown for subjects 3, 16, and 27 in the DISFA dataset. EMOCA and AUGANE capture and reconstruct accurate facial movements within nearby frames. AU 25 (Lips part) of Subject 16 is observable in both DECA and AUGANE, while AU 1 and 2 are observable in EMOCA and AUGANE. Additionally, AUGANE captured the subtle activation changes of AU 1 between adjacent frames of Subject 27. In summary, AUGANE's performance is confirmed to be comparable to state-of-the-art models such as DECA and EMOCA, while surpassing them in certain scenarios.

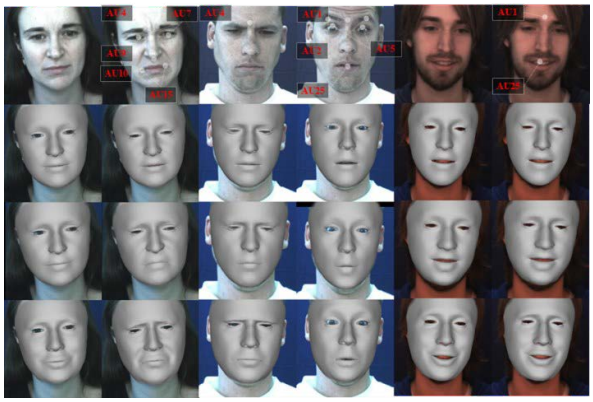


Figure 3 Visual comparison with DECA, EMOCA, and AUGANE on the DISFA dataset. From top to bottom: Input image, DECA, EMOCA, AUGANE.

ACKNOWLEDGMENT

This research was supported by the Kwangwoon University Industry-Academic Collaboration Foundation (2024-2025) and KIAT(Korea Institute for Advancement of Technology) grant funded by the Korea Government (MOTIE : Ministry of Trade Industry and Energy). (P0017124, HRD Program for Industrial Innovation).

REFERENCES

[1] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. AvatarMe: Realistically renderable 3D facial reconstruction "in-the-wild." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 760–769, 2020.

[2] Kristina Scherbaum, Tobias Ritschel, Matthias Hullin, Thorsten Thormählen, Volker Blanz, and Hans-Peter Seidel. Computer-suggested facial makeup. Comput. Graph. Forum, vol. 30, no. 2, pages 485-492, 2011.

[3] Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. Real-time high-fidelity facial performance capture. ACM Trans. Graph., 34(4), jul 2015.

[4] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jae- woo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen- Chun Chen, and Hao Li. Avatar digitization from a single image for real-time rendering. ACM Trans. Graph., 36(6), Nov. 2017.

[5] Diego R. Faria, Mario Vieira, Fernanda C.C. Faria, and Cristiano Premebida. Affective facial expressions recognition for human-robot interaction. Proc. In 26th IEEE Int. Symp. Robot Hum. Interact. Commun. (RO-MAN), pages 805-810, Aug. 2017.

[6] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial an- imation with discrete motion prior. arXiv preprint arXiv:2301.02379, 2023.

[7] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. 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.

[8] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In 2009 sixth IEEE international conference on advanced video and signal based surveillance, pages 296–301. IEEE, 2009.

[9] Y. Feng, H. Feng, M. J. Black, and T. Bolkart. 'Learning an animatable detailed 3D face model from in-the-wild images.' ACM Transactions on Graphics (ToG), Vol.40, No.88, pp.1-13. 2021

[10] Araceli Morales, Gemma Piella, and Federico M. Sukno. Survey on 3d face reconstruction from uncalibrated images. Computer Science Review, 40:1–35, 2021.

[11] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J. Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In Conference on Computer Vision and Pattern Recognition (CVPR), pages 7763–7772, 2019.

[12] Radek Danecek, Michael J. Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and an- imation. In Conference on Computer Vision and Pattern Recognition (CVPR), pages 20311–20322, 2022.

[13] Panagiotis P Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Spectre: Visual speech-informed perceptual 3d facial expression reconstruction from videos. In Proceedings of the IEEE/CVF Conference on CVPR, pages 5744–5754, 2023.

[14] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7680–7689, 2021.

[15] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. arXiv preprint arXiv:2205.01782, 2022.

[16] Paul Ekman and Wallace V. Friesen. Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, 1978.

[17] A. Tewari, M. Zollh¨ofer, H. Kim, P. Garrido, F. Bernard, P. P´erez, C. Theobalt, MoFA: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction, in: Proc. of IEEE ICCV, 2017, pp. 1274–1283.

[18] Brais Martinez, Michel F. Valstar, Bihan Jiang, and Maja Pantic. Automatic Analysis of Facial Actions: A Survey. in IEEE Transactions on Affective Computing, vol. 10, no. 3, pp. 325-347, 1 July-Sept. 2019.

[19] Cheng-Hao Tu, Chih-Yuan Yang, and Jane Yung-jen Hsu. IdenNet: Identity-aware facial action unit detection. In 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), pages 1–8. IEEE, 2019.

[20] Tengfei Song, Lisha Chen, Wenming Zheng, and Qiang Ji. Uncertain Graph Neural Networks for Facial Action Unit Detection. In Proceedings of the AAAI Conference on Artificial Intelligence, 35(7), 5993-6001, 2021.

[21] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. Image and Vision Com- puting, 32(10):692–706, 2014.

[22] S. Mohammad Mavadati, Mohammad H. Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F. Cohn. Disfa: A sponta- neous facial action intensity database. IEEE Transactions on Affective Computing, 4(2):151–160, 2013.

[23] Tengfei Song, Zijun Cui, Wenming Zheng, and Qiang Ji. Hybrid message passing with performance-driven structures for facial action unit detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6267–6276, 2021.

[24] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. arXiv:1710.10903, 2017. PAMI, 23(2):97–115, 2001.

[25] Z. Sun and G. Tzimiropoulos, "Part-based face recognition with vision transformers", Proc. Brit. Mach. Vis. Conf., pp. 1-15, Nov. 2022.

[26] Shangfei Wang, Yanan Chang, and Can Wang. Dual learning for joint facial landmark detection and action unit recognition. IEEE Transactions on Affective Computing, 2021.

[27] Yury Kartynnik, Artsiom Ablavatski, Ivan Gr- ishchenko, and Matthias Grundmann. Real-time facial surface geometry from monocular video on mobile GPUs. In Third Workshop on Computer Vision for AR/VR, Long Beach, CA, 2019.

[28] Nazil Perveen and Chalavadi Krishna Mohan. Configural representation of facial action units for spontaneous facial expression recognition in the wild. In VISIGRAPP (4: VISAPP), pages 93–102, 2020.

[29] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In International Conference on Automatic Face & Gesture Recognition (FG), pages 67–74, 2018.

[30] Dimitrios Kollias and Stefanos Zafeiriou. Aff-Wild2: Ex- tending the Aff-Wild database for affect recognition. arXiv preprint arXiv: 1811.07770, 2018.

[31] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. Retrieved August, 15:2018, 2018.

[32] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4401–4410, 2019.

[33] Yaobin Zhang and Weihong Deng. Class-balanced training for deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 824–825, 2020.

[34] Zhen-Hua Feng, Patrik Huber, Josef Kittler, Peter Han- cock, Xiao-Jun Wu, Qijun Zhao, Paul Koppen, and Matthias R¨atsch. Evaluation of dense 3D reconstruction from 2D face images in the wild. In International Conference on Automatic Face & Gesture Recognition (FG), pages 780–786, 2018