3D Facial Tracking and User Authentication through Lightweight Single-ear Biosensors

Yi Wu, Xiande Zhang, Tianhao Wu, Bing Zhou, Phuc Nguyen, Jian Liu

Abstract—Over the last decade, facial landmark tracking and 3D reconstruction have gained considerable attention due to their numerous applications such as human-computer interactions, facial expression analysis, and emotion recognition, etc. Traditional approaches require users to be confined to a particular location and face a camera under constrained recording conditions (e.g., without occlusions and under good lighting conditions). This highly restricted setting prevents them from being deployed in many application scenarios involving human motions. In this paper, we propose the first single-earpiece lightweight biosensing system, *BioFace-3D*, that can unobtrusively, continuously, and reliably sense the entire facial movements, track 2D facial landmarks, and further render 3D facial animations. Our single-earpiece biosensing system takes advantage of the cross-modal transfer learning model to transfer the knowledge embodied in a *high-grade* visual facial landmark detection model to the *low-grade* biosignal domain. After training, our *BioFace-3D* can directly perform continuous 3D facial reconstruction from the biosignals, without any visual input. Additionally, by utilizing the identical array of ear-worn biosensors, we also showcase the potential for capturing both behavioral aspects, such as facial gestures, and distinctive individual physiological traits, establishing a comprehensive two-factor authentication/identification framework. Extensive experiments involving 16 participants under various settings demonstrate that *BioFace-3D* can accurately track 53 major facial landmarks with only 1.85 mm average error and 3.38% normalized mean error, which is comparable with most state-of-the-art camera-based solutions. The rendered 3D facial animations, which are in consistency with the real human facial movements, also validate the system's capability in continuous 3D facial reconstruction. Experiments also show that the system can authenticate users with high accuracy (e.g., over 99.8% within two trials for three gestures in series), low false positive rate (e.g., less 0.24%), and is robust to various types of attacks.

Index Terms—Mobile computing, wearable sensing, 3D facial reconstruction, user authentication, single-ear biosensing

1 Introduction

Serving as a major role in human interactions, the face conveys both verbal and non-verbal information, such as intention, engagement, and emotion. Facial landmark tracking and 3D reconstruction thus have been becoming fundamental in various emerging applications which require facial analysis. For instance, facial landmark tracking can be used for driver attentiveness monitoring to detect drowsiness and abnormal behaviors [1]. Continuous 3D facial reconstruction can enable a fully immersive user experience by increasing the awareness of the user's real-time facial expressions and emotional states in virtual reality (VR) scenarios [2]. Moreover, recognizing facial movements can enable silent-speech interfaces for convenient human-computer interactions [3]. Additionally, incorporating user authentication alongside facial tracking has the potential to create a more personalized, convenient, and secure user experience. User authentication is essential for numerous privacy-sensitive VR applications (e.g., virtual banking), while identifying the specific user could help to provide personalized experiences based on distinct user preferences. Different from conventional VR authentication schemes in which users are required to input passwords utilizing handheld controllers which is not only inconvenient but also vulnerable to potential side-channel attacks [4], [5], an authentication system capable of verifying the user's identity through simply performing facial expressions would significantly enhance convenience.

Prior Research on Facial Landmark Tracking. Traditional vision-based approaches (e.g., [6], [7], [8]) can localize facial landmarks and produce high-quality facial animations, however, they require a camera positioned in front of the user's face and constrained recording conditions, such as requiring an entire view of the face without occlusions and in good lighting environments. Additionally, a lot of wearable-sensor-based methods have been proposed to recognize user's facial gestures, such as magnetic sensing [9], capacitive sensing [10], and electromyography (EMG)-based sensing [11], [12]. However, all these studies can only distinguish a small set of predefined facial gestures. To the best of our knowledge, there has been no prior work that can continuously track the positions of facial major landmarks (e.g., the mouth, nose, eyes, and eyebrows) and reconstruct 3D facial animations using camera-free and unobtrusive wearable technology.

Prior Research on Wearable/VR Authentication. Commercial EEG headsets (e.g., Emotiv Epoch+ [13]) have been demonstrated to be proficient in distinguishing between various users by harnessing EEG signals [14], [15], [16]. Alternatively, there has been active research on VR authentication leveraging various types of biometrics, including

- *Y. Wu, X. Zhang, T. Wu, and J. Liu are with the Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN, 37996. E-mail: {ywu83, xzhan123, twu21}@vols.utk.edu, jliu@utk.edu;*
- *B. Zhou is with Snap Research. Email: bzhou@snapchat.com*
- *P. Nguyen is with the Manning College of Information & Computer Sciences, University of Massachusetts Amherst, Amherst, MA 01003. Email: vp.nguyen@cs.umass.edu*

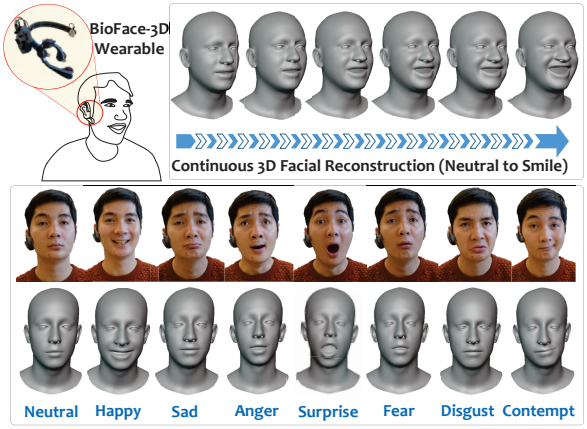


Fig. 1. Illustration of the reconstructed 3D facial avatar with various facial expressions[2].

head motion [17], [18], body motion [19], and ultrasonic reverberations caused by the shape of the users' heads [20]. Nonetheless, none of these studies have the capacity to concurrently conduct 3D facial tracking, and their sensor placement tends to be rather obtrusive (e.g., involving over 10 sensing channels).

System Objective and Challenges. To circumvent all the limitations of existing approaches, this paper aims to provide a wearable biosensing system that incorporates two major functionalities: *(a) Facial Tracking:* Unobtrusively, continuously, and reliably sense the entire facial movements, track 2D facial landmarks, and further render 3D facial animations through fitting a 3D head model to the 2D facial landmarks; and *(b) User Authentication/Identification:* Distinguish different users and authenticate legitimate users through executing a series of pre-defined facial expressions. Although existing studies (e.g., [12], [14], [21]) have shown the success of using biosensors, such as EMG and electrooculography (EOG), to detect facial muscle activities, eye movements, and authenticate users, realizing such a system is still very challenging:

(1) *Biosensing-based Facial Landmark Tracking:* Tracking facial landmarks via biosensing is an unexplored area. Although the captured biosignals can potentially sense expressive facial deformations, it remains unclear how to learn the spatial mapping between the biosignals and facial landmarks.

(2) *Unobtrusive Facial Sensing:* To allow a long-term facial sensing with minimal impact on the user's mobility and comfort, the obtrusiveness and social awkwardness caused by our designed wearable device should be minimized.

(3) *Continuous 3D Facial Reconstruction:* A compelling 3D facial avatar animation requires the rendered 3D faces to be continuous and smooth over time, and the animation should be generated in a timely manner for real-time applications.

(4) *Biosensing-based Authentication/Identification:* Biosignals exhibit significant variability even within the same user, due to variations in gesture intensity, biosensor placement, and shifts in bodily conditions. Therefore, the designed authentication/identification system must possess resilience against these contextual factors.

System Design. To address these challenges, we explore a novel point in the design space and propose a single-earpiece biosensing system, as illustrated in Fig. 1. Specifically, our customized sensing prototype uses two-channel biosensors (i.e., surface electrodes) attached to a very small area around one side of the user's ear to capture both EMG and EOG bioelectrical signals. This sensor position ensures the sensing capability of the biosensing system in providing sufficient information for the entire facial reconstruction while still remaining a minimized obtrusiveness level to the wearer. To enable 3D facial reconstruction beyond the confines of cameras, we build a cross-modal transfer learning model that can learn vision-biosignal correspondences in a supervised manner, which pushes the limits of biosensing to enable rich sensing capabilities that are currently infeasible. More specifically, our designed transfer learning model consists of a visual landmark detection network and a biosignal neural network, enabling facial landmark detection knowledge to be transferred across modalities during training time. During testing, the well-trained biosignal network can directly localize 2D facial landmarks from the biosignals, without any visual input. The recognized 2D facial landmarks will be further processed with a Kalman filter and fitted into a generalized 3D head model to render continuous 3D facial animations. Additionally, due to the variance in signal strength, response, and sensitivity of biosignals on different individuals, we currently adopt user-specific training for BioFace-3D, in which the cross-model mapping is user-dependent. Furthermore, to enable user authentication and identification, we design a CNN-LSTM-based framework with channel- and spatial-wise attention to extract user-specific features from long-term biosignals. The feature representations are further fed into different classifiers for user authentication and identification, respectively. Our main contributions are summarized as follows:

- To the best of our knowledge, *BioFace-3D* is the first single-earpiece biosensing system that can unobtrusively, continuously, and reliably sense the entire facial movements, track 2D facial landmarks, and further render 3D facial animations through fitting a 3D head model to the 2D facial landmarks. The advanced biosensing system also demonstrates potential for two-factor user authentication and identification based on biosignal.
- Through a thorough anatomical analysis of human facial muscles and elaborate experiments, we identify optimal biosensor placement positions on the face to maintain a minimized obtrusiveness level of the sensing prototype.
- Relying on the transfer learning across multiple modalities, we push the limits of biosensing to make it possess the capability of other *high-grade* modalities (e.g., vision). This significantly extends its sensing capabilities beyond the common form of biosensing and introduces new opportunities for many emerging applications.
- Extensive experiments involving 16 participants and various settings demonstrated the effectiveness and robustness of the system. The results show that *BioFace-3D* can accurately track 53 facial landmarks with only

2. Our rendered facial animation samples can be found at https://mosis.eecs.utk.edu/bioface-3d.html.

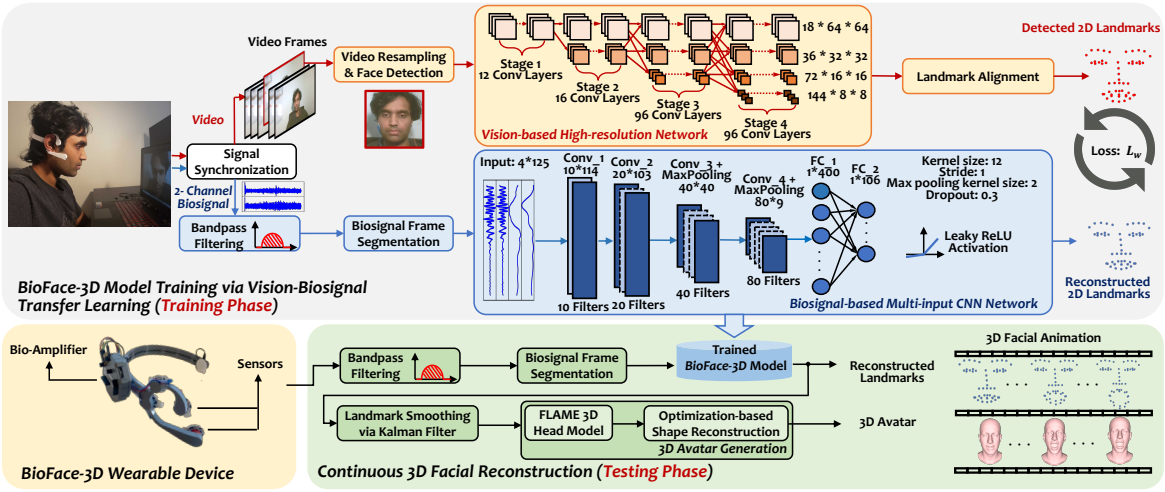


Fig. 2. *BioFace-3D* facial tracking system overview.

1.85 mm average error and 3.38% normalized mean error, which is comparable with most camera-based solutions. The system additionally boasts remarkable performance in user authentication, achieving a true positive rate surpassing 99.8% within two trials for three gestures in series, accompanied by a notably low false positive rate (e.g., below 0.24%).

Our preliminary work has been published in ACM MobiCom 2021 [22]. In this journal paper, we present extensive revisions and enhancements centered around harnessing biosignals for user authentication/identification. Additionally, we elevate the standard 3D avatar concept by fashioning user-specific, photorealistic animations. These animations incorporate intricate facial details, resulting in a heightened sense of immersion and personalization.

2 System Overview

As shown in Fig. 2, the proposed *BioFace-3D* has two phases: the *training phase* in which our system uses the biosignals and visual information in a supervised manner to learn the real-time behavioral mapping from biosignal stream to facial landmarks, and the *testing phase* where the well-trained biosignal network can work independently to perform continuous 3D facial reconstruction, without any visual input. Specifically, during training, we collect visual and biosignal streams using an off-the-shelf camera (e.g., a laptop's built-in camera) and our designed *BioFace-3D* wearable device (Appendix C), respectively. We then perform *Signal Synchronization* to ensure the synchronization between the streamed biosignal and the video frames. After that, the visual and biosignal streams are separately processed as follows:

Visual Stream in Training. We first conduct *Video Resampling* to make the recorded videos from different camera types to be resampled in a uniform frame rate, which allows the vision network to take any visual input regardless of its actual frame rate in recording. Next, we perform *Face Detection* for each video frame, and crop the frame to only preserve the detected face. The cropped image frames are then fed into the pre-trained *Vision-based High-resolution Network* for 2D facial landmarks detection. Furthermore, we employ *Landmark Alignment* to eliminate the effect caused by head poses (i.e., scale, rotation, and translation). The detected 2D facial landmarks are then warped and transformed into a uniformly aligned coordinate space, which will serve as the ground truth to guide the training of the biosignal network. Please note that the choice of the vision-based model can be adjusted to suit the particular demands of the applications. Section 4.2 introduces how to reconstruct 3DMM parameters and create a more personalized photo-realistic animation.

Biosignal Stream in Training. *BioFace-3D* collects two biosignal streams from the biosensors integrated into our single earpiece wearable. Each biosignal stream is first processed to obtain both EOG and EMG biosignal streams via *Bandpass Filtering* [23]. We then apply *Biosignal Frame Segmentation* to segment the filtered biosignal stream into frames, each corresponding to a re-sampled video frame. The signal segments are then fed into *Biosignal-based Multi-input CNN Network* to reconstruct 2D facial landmarks. To transfer knowledge from the vision network into the biosignal domain, we utilize the Wing loss [24] to enhance attention of the landmarks which are important but less active (e.g., pupils) and to help the biosignal network learn an accurate spatial mapping between biosignals and facial landmarks.

Biosignal Stream in Testing to Continuously Reconstruct 3D Faces. During testing, the biosignal stream first passes through the same pre-processing procedures in training. Then the fine-tuned biosignal network can continuously reconstruct 2D facial landmarks from the biosignal stream, without any visual input. To ensure a fluent 3D avatar animation, we then apply *Landmark Smoothing via Kalman Filter* to stabilize the facial landmark movement across successive frames. Next, we generate 3D facial animation from the stabilized landmarks using the FLAME (Faces Learned with an Articulated Model and Expressions) model [25]. The generated sequence of fitted head models can then be used for rendering a 3D facial animation that recovers the user's facial movements.

3 Biosignal-based Facial Landmark Reconstruction via Knowledge Transfer

In this section, we describe the detailed training procedure and the designed knowledge transfer learning network across multiple sensing modalities.

3.1 Signal Synchronization

To guarantee the synchronization between the two modalities' data streams, the user needs to tap the earpiece near the bottom measurement sensor at the beginning of the training phase. This way, a sharp and sizeable peak will be generated in the biosignal stream due to the *skin-electrode contact variation*, while such an event can also be tracked in the video stream with quantifiable accuracy (e.g., through detecting the user's hand using a pre-trained hand keypoint detection model [26]). To detect such a peak in the biosignal stream, we implement a z-score peak transformation algorithm [27], which calculates if any data point of the biosignal stream deviates from a moving average by a given threshold τ. In our implementation, we use a moving window size of 40 milliseconds across all users, which is sufficient to detect the signal peak caused by the finger tap. The threshold τ is set to $\mu_w \pm 0.4\sigma_w$, where μ_w and σ_w are the mean and standard deviation of the sliding window. This z-score based method has been shown to be effective and accurate throughout our system evaluation.

3.2 Data Pre-processing

Visual Stream - Video Resampling & Face Detection. To make our system compatible with various recording devices of different frame rates, we first downsample the recorded video to a uniform frame rate $f_v = 20$, which can also reduce the computational cost for real-time facial reconstruction while maintaining the fluency of the video. Specifically, given the frame rate of the original video f_o, we only keep $\frac{f_v}{f_o}$ of the frames equally distributed in the video buffer, and the timestamps of these frames are then re-scaled to the new timebase (i.e., $\frac{1}{f_v}$). After resampling, we apply a pre-trained Haar Cascade Classifier [28], which provides high accuracy in object detection under varied lighting conditions, to each downsampled video frame for face detection. To meet the required input size of the following vision network, we then make the detected face centered, crop the corresponding square area, and resize the cropped frame to 256×256 pixels.

Biosignal Stream - Bandpass Filtering & Biosignal Frame Segmentation. On the biosignal side, we first apply two band-pass filters to extract the main structure of the bio-electrical signals, i.e., EMG and EOG bioelectrical signals [23]. Moreover, in order to transfer knowledge from the vision-based facial landmark detection model into the biosignal modality, we need to match the visual input (i.e., resampled video frames) with the time-series biosignal input. To match with each video frame, we segment the biosignal streams (i.e., both EMG and EOG signals) into overlapped short frames starting at each video frame's timestamp. Given that the gap between adjacent frames is $\frac{1}{f_v} = 0.05s$ and the sampling rate of biosignal is 250 Hz, the length of each biosignal frame is set to $l = 0.5s$ for all experiments, which creates massive overlapped data samples between adjacent biosignal frames as well as sufficient data for the CNN network. This setting makes the subsequent transfer learning model better capture the temporal dynamics and dependencies among continuous biosignal streams to ensure smooth frame transitions in the rendered animation.

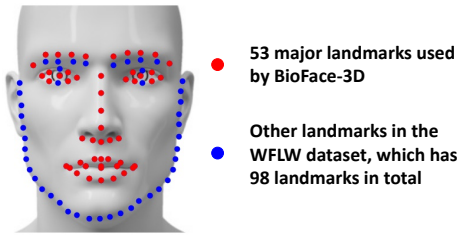


Fig. 3. Major facial landmarks used in *BioFace-3D*.

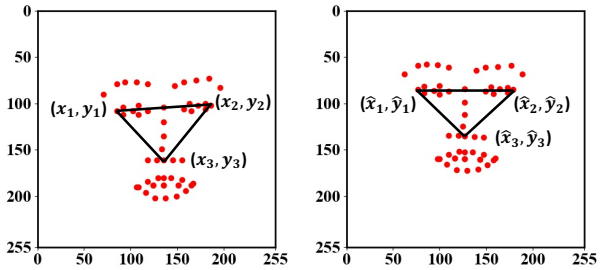
3.3 Vision-based High-resolution Network

Conventional image-processing networks for facial landmark detection either rely on low-resolution features built by gradually reducing the size of the feature maps (e.g., TCNN [29]), or utilize a 2-stage high-to-low and low-to-high process to first extract low-resolution features and then rebuild high resolution features through deconvolution and unpooling operations (e.g., encoder-decoder [30]). However, the important spatial and semantic information embedded in the initial high-resolution features might be lost during this process and is hard to recover. To address this and improve the recognition accuracy, we adopt a high-resolution network (HRNet) [8] which maintains high-resolution through the whole process. As shown in Fig. 2, the whole network consists of four stages, in which low-resolution convolution streams are added gradually during the training process.

Specifically, the first stage only has a single high-resolution (64×64) stream with 12 convolutional layers, and the depth is set to 18. The subsequent stages decrease the resolution to $\frac{1}{2}$ of the resolution of the previous stage and double its depth. Stage 2 adds a lower resolution stream and the number of layers is increased to 16, while Stage 3 and Stage 4 handle more streams in parallel using 96 convolutional layers, with 16×16 and 8×8 resolution, respectively. Each stage processes a number of convolution streams with different resolutions in parallel. At the end of each stage, information is exchanged among different resolutions via repeated multi-resolution fusions, where low-resolution representations are up-sampled and concatenated with the high-resolution representation. Specifically, we use a pre-trained model on the WFLW dataset [31], which has a total of 98 landmarks, as shown in Fig. 3. To reduce computational complexity, we only keep 53 landmarks that cover major facial components such as eyes, eyebrows, nose, and mouth. The output of the vision-based facial landmark detection network provides biosignal modality with transferable knowledge for training the biosignal network.

3.4 Landmark Alignment

The detected landmark positions can be impacted by large head pose variations caused by head motions, facing directions, and camera angles and positions. To eliminate the impact of these irrelevant factors, we attempt to obtain a canonical alignment of the face based on affine transformations including translation, rotation, and scaling. Specifically, given the coordinate of the i_{th} facial landmark (x_i, y_i), the transformed landmark $(\hat{x_i}, \hat{y_i})$ can be obtained by:



(a) Before landmark alignment (b) After landmark alignment

Fig. 4. Illustration of facial landmark alignment.

$$\begin{bmatrix} \hat{x_i} \\ \hat{y_i} \\ 1 \end{bmatrix} = \mathbf{R} \cdot \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}, \quad (1)$$

where $\mathbf{R}$ is the affine matrix. To derive $\mathbf{R}$, we fix the positions of three aligned landmarks (i.e., the left canthus $(\hat{x_1}, \hat{y_1})$, the right canthus $(\hat{x_3}, \hat{y_3})$, and the tip of the nose $(\hat{x_2}, \hat{y_2})$) which are supposed to be static in the aligned coordinate space, as shown in Fig. 4. To be more specific, given the video frame size of $w \times w$, the coordinates of two lateral canthus are fixed to $(\lfloor \frac{3w}{10} \rceil, \lfloor \frac{w}{3} \rceil)$ and $(\lfloor \frac{7w}{10} \rceil, \lfloor \frac{w}{3} \rceil)$, respectively. According to ideal facial proportions [32], the tip of the nose is fixed to $(\lfloor \frac{w}{2} \rceil, \lfloor \frac{8w}{15} \rceil)$. With the three fixed landmarks' coordinates and the coordinates before alignment, we can derive all the unknown entries in $\mathbf{R}$ through solving a set of six-variable linear equations. We can then use Equation 1 to align all the remaining facial landmarks.

3.5 Biosignal-based CNN Network

During the training of the biosignal network, we take the aligned 2D facial landmarks from the vision network as ground truth and train a 1D CNN network to regress the facial landmarks directly from four channels of time-series biosignals (i.e., two EMG and two EOG streams). Other network architectures (e.g., TDNN and LSTM) may also work, but 1D CNN is more suitable for end-to-end learning of raw time-series data and has relatively lower computational cost [33]. Specifically, given the default sampling rate of the biosignal f_s = 250 Hz and the length of each biosignal frame l = 0.5 s, the input size of the biosignal network is 4×125. The output of the network is the 2D coordinates of 53 facial landmarks. As shown in Fig. 2, the network has 4 1D convolutional layers and 2 fully-connected layers. Each convolutional layer has a kernel length of $\lfloor \frac{f_s}{f_v} \rceil$), which is the time gap between adjacent frames. Additionally, the number of filters is doubled when the network is processed to the subsequent convolutional layer, which is initially set to 10. Two max-pooling layers are added to the last two convolutional layers to obtain a more compressed feature.

Loss Function. Training our learning model comes down to minimizing the designed loss functions to decrease the error between predicted landmark positions and the corresponding ground truths. As landmark position loss treats each individual landmark independently, some important but less-active landmarks, such as pupils compared with lips, may not achieve good attention during training because all the landmarks share an equal weight. To address this issue, we adopt the wing loss function [24], and the loss for each facial landmark is defined as:

$$Loss(x_i) = \begin{cases} w * ln(1 + |x_i|/\epsilon), & \text{if } |x_i| < w \\ |x_i| - C, & \text{otherwise.} \end{cases} \quad (2)$$

where $|x_i|$ is the L2 distance between the ground truth and the reconstructed coordinate for the i_{th} landmark. w represents the threshold of the small error, which is set to 20 in our case. ϵ means the curvature in the small error range, and $C = w - wln(1 + w/\epsilon)$ which links the linear part and non-linear part together. This way the small range errors would obtain more attention when training a regression network, thereby significantly improving the network training capability for the small-scale error landmarks.

Optimization. In addition, the network is trained using the Adam optimizer [34], and the learning rate is set to 0.1 with a decay of 0.9 every 10 epochs. The stride and dilation are all set to 1, and each layer has a dropout rate of 0.3 to avoid over-fitting.

4 Continuous 3D Facial Reconstruction

In this section, we mainly introduce the testing phase of *BioFace-3D*. Specifically, the well-trained biosignal network takes as input each pre-processed biosignal frame to reconstruct 2D facial landmarks. Then, a Kalman filter and a 3D head model are used to stabilize landmarks and generate 3D facial animation, respectively.

4.1 Landmark Smoothing via Kalman Filter

We observe that the reconstructed facial landmarks regressed directly from the biosignal network are inevitably jittery, which may be caused by the instability of the network as well as the noises introduced in the biosignal. To guarantee the smoothness of the reconstructed 2D facial landmarks over time, we adopt a Kalman filter [35] to stabilize the landmark outputs. Specifically, given a facial landmark in the frame t, we define its state vector $\mathbf{s_t} = [x^t, y^t, v_x^t, v_y^t, a_x^t, a_y^t]^T$, where x^t, v_x^t, a_x^t represents the location, velocity, and acceleration of the landmark, respectively, along x axis, while y^t, v_y^t, a_y^t stands for y axis. A state-space model describing this landmark movement thus can be represented as $\mathbf{s_t} = \mathbf{A}\mathbf{s_{t-1}}$, and the landmark coordinates $\mathbf{z_t} = \mathbf{H}\mathbf{s_t}$, where the state transition matrix $\mathbf{A}$ and the observation matrix $\mathbf{H}$ can be defined as:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & \Delta t & 0 & \frac{1}{2}\Delta t^2 & 0 \\ 0 & 1 & 0 & \Delta t & 0 & \frac{1}{2}\Delta t^2 \\ 0 & 0 & 1 & 0 & \Delta t & 0 \\ 0 & 0 & 0 & 1 & 0 & \Delta t \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \mathbf{H} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}^T, \quad (3)$$

where $\Delta t = \frac{1}{f_v}$ represents the time interval between two adjacent frames. Given the known constant variable Δt and the frame size of 256×256, based on the relationships between the six variables in $\mathbf{s_t}$, the process and measurement noise covariances, $\mathbf{Q}$ and $\mathbf{R}$, are set to:

$$\mathbf{Q} = \begin{bmatrix} \frac{1}{4}\Delta t^4 & 0 & \frac{1}{2}\Delta t^3 & 0 & \frac{1}{2}\Delta t^2 & 0 \\ 0 & \frac{1}{4}\Delta t^4 & 0 & \frac{1}{2}\Delta t^3 & 0 & \frac{1}{2}\Delta t^2 \\ \frac{1}{2}\Delta t^3 & 0 & \Delta t^2 & 0 & \Delta t & 0 \\ 0 & \frac{1}{2}\Delta t^3 & 0 & \Delta t^2 & 0 & \Delta t \\ \frac{1}{2}\Delta t^2 & 0 & \Delta t & 0 & 1 & 0 \\ 0 & \frac{1}{2}\Delta t^2 & 0 & \Delta t & 0 & 1 \end{bmatrix}, \mathbf{R} = \begin{bmatrix} 12.5 & 0 \\ 0 & 12.5 \end{bmatrix}. \quad (4)$$

The process noise covariance is a covariance matrix associated with the errors in the state vector s_t, where the noise of acceleration is initialized to 1. This covariance will automatically get updated to achieve a good state. The values in the measurement noise covariance matrix are set to a relatively large value, which ensures that jittery landmarks with larger errors can still be effectively smoothed. The smoothed landmark coordinate in the frame t can then be derived as $\mathbf{H}\hat{\mathbf{s}}_\mathbf{t}$, where $\hat{\mathbf{s}}_\mathbf{t}$ is the optimal state estimate.

We calculate the average standard deviation of all mouth-related landmarks as the evaluation metric to validate the effectiveness of the Kalman filter. Specifically, we select 4 minutes of reconstructed landmarks in which the user repeatedly performs the *surprise* expression. In addition to the Kalman filter, we also implement a simple linear interpolation technique, in which the average of each adjacent frame pair is compensated between them. Specifically, the average standard deviation of all mouth-related landmarks is 8.66 if no smoothing techniques are applied, 8.61 when simple linear interpolation is utilized, and 7.73 when the Kalman filter is implemented. The results demonstrate the effectiveness of the Kalman filter on landmark smoothing.

4.2 3D Avatar Generation

To improve system usability and reduce modeling complexity, we seek a compact head model that can be easily fitted to data while preserving enough details to generate expressive facial animations.

FLAME 3D Head Model. The FLAME (Faces Learned with an Articulated Model and Expressions) model [25] is a statistical 3D head model that uses a learned shape space of identity variation and articulated jaw, neck, and eyeballs to achieve accurate, expressive, and computationally efficient 3D face modeling. The model is based on linear blend skinning and corrective blendshapes, and contains 5023 vertices and 4 rotary joints (neck, jaw, and eyeballs). The modeling process can be viewed as a function: $M(\vec{\beta},\vec{\theta},\vec{\psi}) : \mathbb{R}^{|\vec{\beta}|\times|\vec{\theta}|\times|\vec{\psi}|} \rightarrow \mathbb{R}^{3N}$, that takes shape $\vec{\beta} \in \mathbb{R}^{|\beta|}$, pose $\vec{\theta} \in \mathbb{R}^{|\theta|}$, and expression coefficients $\vec{\psi} \in \mathbb{R}^{|\psi|}$ and return N vertices. The model is composed of a template mesh of a neutral pose, shape blendshapes, pose blendshapes, and expression blendshapes, which are used to account for variations caused by identity, pose deformation, and facial expressions, respectively.

Optimization-based Shape Reconstruction. To generate a 3D head model that reflects the user's facial movements and expressions, we exploit a 2-stage optimization process to fit the generic 3D head model to the 2D landmarks extracted from biosignals. In the first stage, we conduct camera calibration by optimizing the parameters for rigid transformation, including scale, rotation, and translation, to minimize the L_2 distance between the landmarks and the corresponding 3D head model vertices projected into the 2D space. In the second stage, we optimize the model parameters (e.g., pose, shape, and expression) by optimizing the L_2 distance while regularizing the shape coefficients $\vec{\beta}$, pose coefficients $\vec{\theta}$ (including neck, jaw, and eyeballs), and expression coefficients $\vec{\psi}$ by penalizing their L_2 norms. After optimization, we can generate a 3D head model that recovers the user's facial expressions.

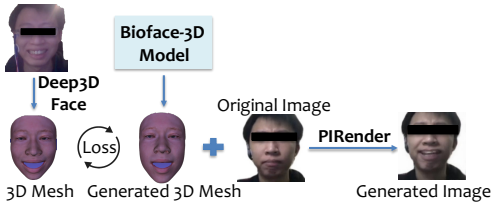


Fig. 5. Overview of photo-realistic animation generation.

4.3 Photo-realistic Animation Generation

FLAME utilizes a generic head model that lacks detailed features and awareness of the user's face. Consequently, we provide users with an additional option to generate personalized, photo-realistic animations that incorporate specific facial details instead of the generic 3D animation. Specifically, instead of 2D facial landmarks, we change the output of the biosignal-based CNN network to 3D face meshes, which are represented using 3D Morphable Model (3DMM) [36] coefficients. As depicted in Fig. 5, different from sparse 2D landmarks, 3D face meshes capture a more densely detailed geometry and preserve personalized facial features. We use Deep3DFace [37] to extract 3DMM coefficients from facial images as the groundtruth. Deep3DFace could be considered as a modified ResNet-50 [38] network, where the size of the last fully-connected layer has been adapted to 239. This layer is structured to represent 3DMM coefficients representing identity, expression, texture, pose, and lighting of the input facial image. We change the final fully-connected layer of the biosignal-based 1DCNN network to 239, and use L2 loss to reconstruct the 3DMM coefficients. We set the learning rate to 0.001 with a decay of 0.95 every 10 epoches using the Adam optimizer. The dropout rate is 0.3.

As illustrated in Fig. 5, the reconstructed 3DMM coefficients are further combined with an arbitrary photo of the user to synthesize the photo-realistic animation. We utilize Portrait Image Neural Renderer (PIRenderer) [39], another pre-trained deep learning model that utilizes 3DMM coefficients to manipulate facial expressions and motions in arbitrary facial images. PIRender is composed of three sub-networks: a mapping network that maps the 3DMM coefficients to a latent vector; a warping network that estimates the difference between the input facial image and the desired facial expressions based on the latent vector, and generates coarse results through wrapping the input image with the estimated deformations; and an editing network which refines the coarse results and produces the final photo-realistic images. The generated image in Fig. 5 captures the intended facial expression while retaining the individual's personalized facial details.

5 User Authentication & Identification via BioFace-3D

5.1 Threat Model

We consider the following two application scenarios that require to distinguish users: *(1) User Authentication:* In this scenario, our aim is to authenticate the identity of a sole legitimate user, granting access to a security-sensitive service, while simultaneously denying entry to any other individuals attempting to use the VR system; and *(2)*

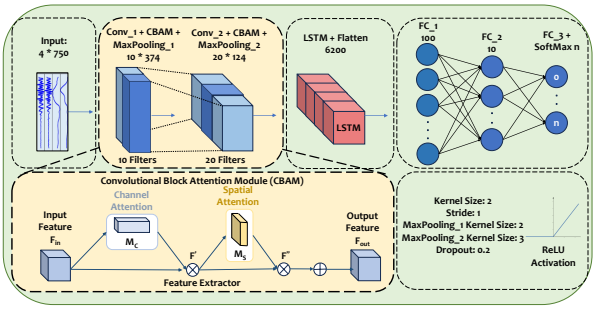


Fig. 6. Architecture of user authentication & identification.

User Identification: We consider a scenario in which a VR device is shared among a group of individuals, and our goal is to distinguish the unique user to offer customized experiences. To demonstrate the reliability of the proposed system, we consider the following attacks that are harmful to the proposed authentication functionalities:

Blind Attack. A potential adversary seeks to circumvent the authentication system or masquerade as a specific user for user identification by attempting random facial expressions while wearing the wearable prototype.

Credential-aware Attack. The adversary possesses knowledge of the authorized user's credentials, which include a prescribed sequence (consisting of three to five) of facial expressions. With this information, the adversary endeavors to imitate the genuine user's facial expressions in an attempt to deceive the user authentication/identification system.

5.2 Data Augmentation

To address the inherent class imbalance commonly encountered in user authentication datasets, we implemented data augmentation techniques exclusively within the training dataset, enhancing the resilience of our models. Our approach consists of a dual strategy aimed at rectifying the skewed class distribution. Initially, the underrepresented class (e.g., legitimate user) is amplified using the Synthetic Minority Over-sampling Technique (SMOTE) algorithm [40], which generates synthetic samples, allowing us to achieve a more equitable representation of classes. This technique enables precise adjustment of the minority-to-majority class ratio, fine-tuning the upsampling process. Furthermore, we apply additional data augmentation through signal-based modifications, encompassing random time shifts of up to 10 ms, injection of Gaussian noise, and amplitude scaling variations of up to 10% to further increase the diversity of the datasets. This strategy effectively mitigates class imbalance, enhancing the model's adaptability to diverse scenarios.

5.3 System Overview for User Authentication & Identification

Fig. 6 illustrates the deep learning model architecture for user authentication & identification. In the current design, the user is prompted to execute either an individual facial gesture or a sequence of such gestures within a brief temporal window for user verification or identification. Upon completion of each facial gesture input, similar to the pre-processing steps for 3D facial tracking (Fig. 2), the four-channel biosignals are initially filtered through dual band-pass filters to extract EMG and EOG signals correspondingly. The signals will then be fed into a CNN-LSTM hybrid neural network, which can effectively combine feature extraction and time series regression for deep learning and make full use of the spatio-temporal correlation of the biosignals, for identity verification/identification. By treating facial gestures as user-owned passcodes and harnessing the distinctive individual traits encoded within the biosignals, the system facilitates a fortified two-factor framework for user authentication and identification.

To authenticate/verify users, the biosignals first pass through two 1D CNN layers with 10 and 20 filters, respectively. The kernel size is set to 2, with the stride length set to 1. Additionally, we add Convolutional Block Attention Module (CBAM) [41], which infers attention maps along channel and spatial dimensions and assigns weights to more important features. Two max-pooling layers with kernel sizes of 2 and 3, respectively, are utilized after each convolution layer to further down-sample features. Different from the facial tracking task, which requires rapid inference (i.e., 20 FPS) and short-term sequences (i.e., 0.5s), the user authentication/identification task doesn't require such frequent inference, and the input length is significantly longer. We therefore apply an additional Long Short-Term Memory (LSTM) layer [42] after CNN layers to further capture long-term dependencies in the biosignal, which enables the model to effectively extract both spatial and temporal representations from biosignals. Specifically, we set the hidden layer size of the LSTM layer to 50, and the output is flattened and fed into three fully connected layers. The first two layers have 256 and 128 units, respectively, while the size of the last layer is 2 (i.e., legitimate user & adversary) for user verification and N for user identification, where N is the number of enrolled users in the system. We use ReLU as the activation function and apply a dropout rate of 0.2 to avoid overfitting. We utilize cross-entropy as the loss function, and the network is trained using the Adam optimizer with the learning rate set to 0.0005.

5.4 Series of Gestures & Majority Vote

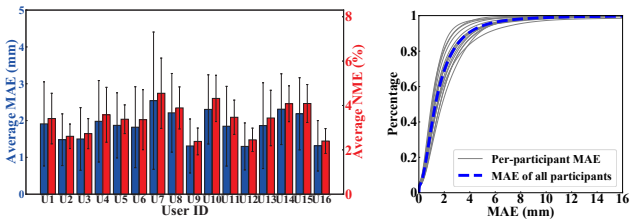
To further enhance system performance, users have the option to execute gestures multiple times, and the authentication/identification network will provide corresponding multiple prediction results. For the authentication network, we utilize a hard majority vote, where the final prediction equals the result generated by the majority of the gestures. For the identification network, to deal with circumstances in which conflicting predictions arise (e.g., an extreme scenario where each of the predictions corresponds to a different individual), we employ a soft majority vote, wherein the softmax output of multiple gestures is averaged to obtain the final result.

6 Performance on Facial Tracking

6.1 Experimental Methodology

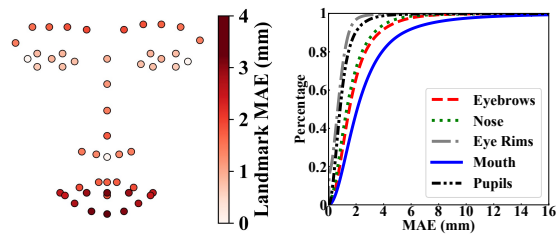
Experimental Setup & Data Collection. We recruited 16 participants to evaluate the performance of *BioFace-3D*[3]. Particularly, the participants include 11 males and 5

3. The study has been approved by our Institutional Review Board (IRB).



(a) Per-participant landmark tracking error (b) MAE CDF for each participant

Fig. 7. Performance of continuous facial landmark tracking for each participant.



(a) Visualization of the average MAE (b) MAE CDF for each facial feature

Fig. 8. Performance of continuous facial landmark tracking for each facial landmark and facial feature.

females, aging from 21 to 34 years old. Six of them wore glasses during the data collection as usual. To evaluate the performance of tracking 53 facial landmarks, we focus on seven universal facial expressions of emotion [43] involving *happy*, *sad*, *anger*, *surprise*, *fear*, *disgust*, and *contempt*, as shown in Fig. 1. The participants were asked to sit in front of a camera (for training and ground truth recording purposes) and repeatedly perform the aforementioned seven expressions while wearing our implemented *BioFace-3D* prototype. Each expression was separated by a *neutral* facial expression (i.e., relaxed facial expression). To assist participants with their data collection, seven pictures were displayed on a screen portraying the corresponding faces for them to imitate. The pace and to which extent each expression was performed were not controlled throughout the experiments. To show the generalizability of our system in using various types of cameras for training, we used a variety of cameras of different resolutions and recording frame rates (e.g., 720P, 1080P resolutions, and 25, 30 fps), including the webcam of a Lenovo ThinkPad X1, a Lenovo Ideapad Y700, a MacBook Pro 2019, an EMeet C960 Webcam on a desktop, and the built-in rear camera of an iPhone 8.

Particularly, each participant was asked to repeatedly make each facial expression for 4 minutes, which leads to about 40 to 50 rounds of facial expressions. The data collection lasts for 28 minutes (7 facial expressions in total) for each participant, and their eye movements were not constrained during the data collection. Unless mentioned otherwise, for each participant, we use the first 20 minutes of data for training and the remaining 8 minutes of data for testing. The default sampling rate of biosignals per channel was set to 250 Hz. The impact of sampling rate on performance will be discussed in Appendix D.3.1. After data collection, we also asked participants to complete a questionnaire on their experience with *BioFace-3D*, which is elaborated in Appendix D.4. We also extended our experiments to other types of facial movements (i.e., speaking) with 5 participants involved, which is detailed in Appendix D.2. We further collected 4 additional datasets with one participant involved to study the impact of facial occlusion and bursty head movements. Three participants separately evaluated the performance of eye tracking and tested the system's temporal stability when training and testing data are separated by multiple days. The data collection details for these tests are elaborated in Appendix D.1 and Appendix D.3.

Evaluation Metrics. *1) Mean Absolute Error (MAE)* is the absolute error between the reconstructed landmarks and groundtruth landmarks, which are converted from pixels to a physical unit (millimeter). The MAE of a single landmark can be calculated as $MAE = ||g - r||_2 \times \frac{l_r}{l_f}$, where g and r represent the groundtruth and reconstructed landmark coordinates, respectively. l_f is the distance between the two lateral canthus in the frame, which is $\lfloor \frac{2w}{5} \rceil$ as aforementioned in Section 3.4, while l_r is the distance between the two lateral canthus of the participant we measured; *2) Normalized Mean Error (NME)* is the mean error between the groundtruth and reconstructed landmark coordinates, normalized by the inter-ocular distance, which is a commonly used metric in camera-based solutions for facial landmark tracking. Given the groundtruth and reconstructed coordinates of landmark as g and r, the NME can be calculated as $NME = \frac{||g-r||_2}{||g_{lp}-g_{rp}||_2}$, where g_{lp} and g_{rp} denote the groundtruth of left pupil and right pupil, respectively.

6.2 Overall System Performance

6.2.1 Facial Landmark Tracking (Facial Expression)

Fig. 7 (a) illustrates the average MAE & NME and corresponding standard deviations for all the 53 facial landmarks of each participant. We observe that all the participants can achieve comparable low errors. Specifically, *BioFace-3D* obtains an average of 1.85 mm MAE and 3.38% NME with average standard deviations of 0.99 mm and 0.90%, respectively, indicating that mm-level accuracy could be achieved in our system. Among all the participants, *U12* achieves the best reconstruction results with only 1.29 mm MAE and 2.45% NME, while *U7* has the largest error (i.e., only 2.54 mm MAE though). Fig. 7 (b) depicts the Cumulative Density Function (CDF) of the MAE errors for each individual participant as well as cross-participant cases. 80% of the reconstructed landmarks have a low MAE of < 2.66 mm, which demonstrates the promising capability of *BioFace-3D* in tracking human 2D facial landmarks.

In addition, distinct landmarks may have different scales of errors due to their movement variability. Fig. 8 (a) visualizes the average MAE for the entire 53 major landmarks, and Fig. 8 (b) shows the CDF of the categorized landmarks. We find that reconstructed landmarks on the mouth have a relatively larger error, but 80% of them are still within an acceptable range (i.e., < 3.87 mm). Eye rims (12 landmarks in total without pupils) and pupils (2 landmarks only) achieve a relatively lower MAE error. Specifically, 80% of the reconstructed eye-related landmarks are within 1.17 mm, indicating *BioFace-3D* can

TABLE 1
Comparison with vision-based solutions.

Methods	Dataset	# of Landmarks	NME
SDM [44]	300-W	68	7.52
	LFPW	68	5.67
CFSS [45]	300-W	68	5.76
	LFPW	68	4.87
HRNet [8]	300-W	68	2.87
	WFLW	98	4.60
BioFace-3D	**Self-collected**	**53**	**3.38**

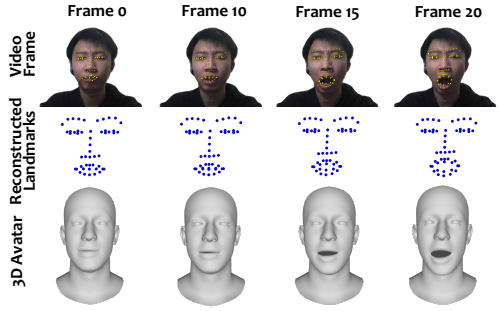


Fig. 9. Example of the rendered facial animation.

accurately track the unconstrained eye movements of the participants during data collection.

As NME is a commonly used metric in vision-based facial landmark tracking, we directly compare our landmark tracking results with several state-of-the-art vision-based solutions [8], [44], [45] in Table 1. These vision-based solutions were evaluated using multiple public image datasets (e.g., WFLW [31], 300-W [46]) which have manually labeled groundtruths and different numbers of facial landmarks to be reconstructed. Although it might not be a fair comparison as our dataset is self-collected and we use a pre-trained camera-based network to generate landmark groundtruths instead of human labeling, the comparable NME accuracy shows the promising performance of *BioFace-3D*, even compared with vision-based solutions.

6.2.2 Continuous 3D Facial Reconstruction

To test *BioFace-3D*'s ability of continuous 3D facial reconstruction, we show the video frames, reconstructed landmarks, and rendered 3D avatar frames at a interval of 5 frames in Fig. 9. Our rendered facial animation samples can be found at [47]. We observe that the final 3D facial animation and 2D landmark generated from the biosignal features closely resemble the animation and landmarks generated using visual features, demonstrating the effectiveness of the biosignal features on capturing facial dynamics.

6.2.3 Photo-realistic Animation Rendering

To demonstrate the versatility of our proposed system across various avatar rendering scenarios, we visualize different facial expressions, reconstructed 3DMM coefficients, and photorealistic synthesized images in Fig. 10. We can find that *BioFace-3D* is able to generate photo-realistic synthesized images representing different facial expressions. The generated photo-realistic avatar videos can also be found at [47].

Fig. 10. Example of photo-realistic animation generation.

7 Performance on User Authentication & Identification

7.1 Experimental Methodology

Experimental Setup & Data Collection. We evaluate the performance of *BioFace-3D* for user authentication/identification using data previously collected from 16 participants (detailed data collection procedures are outlined in Section 6.1). In the authentication scenario, each participant will take turns to be selected as the legitimate user, with all other participants considered as attackers. Data from 9 of these attackers will be employed for training the authentication network, while the data from the remaining 6 will not be used in the network's training. This approach aims to create a more realistic scenario for evaluation. In the blind attack scenario, the attacker will randomly perform a gesture in an attempt to fool the authentication network, while in the credential-aware attack scenario, the attacker will mimic the same gesture as the legitimate user. In the identification scenario, given the use case involving a small group of family members sharing the same device, we randomly selected five participants for training the identification network. Additionally, we extend the evaluation to accommodate up to 16 participants enrolled in the system. For all of the scenarios, we divide the whole dataset into three parts: training data, validation data, and testing data with a ratio of 8:1:1.

Evaluation Metrics. *1) True Positive Rate (TPR)* is the probability of the legitimate user successfully pass the authentication system; *2) False Positive Rate (FPR)* is the rate of attackers passing the system; *3) Receiver Operating Characteristic (ROC) Curve* visualizes the relationship between FPR and TPR under varying threshold settings; *4) Area Under Curve (AUC)* reflects the area underneath the ROC curve, which provides an aggregate measure of system performance across all thresholds, with a higher value indicating better model performance; *5) Identification Accuracy* is the probability of a user being correctly classified; *6) Confusion Matrix* is a tabular representation that illustrates the classification results of the identification network. The shade of the cells in the matrix indicates the proportion of users that are correctly classified.

7.2 Performance on User Authentication

Single User Authentication. Fig. 11 illustrates the TPR & FPR of user authentication across all 16 participants. Specifically, each participant performs three gestures in a row, and we compute the average result based on all potential combinations of gestures. We observe that the

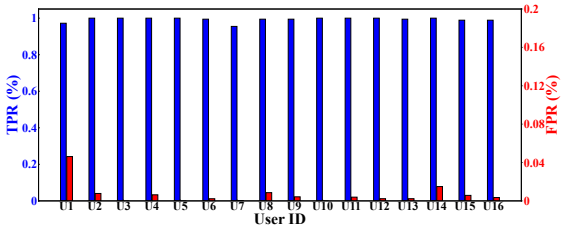
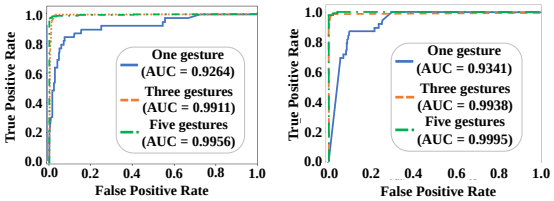


Fig. 11. Average of all combinations for three gestures in series' True Positive Rate and False Positive Rate for each user.



(a) ROC curves of user authentication without using CBAM

(b) ROC curves of user authentication using CBAM

Fig. 12. Performance of user authentication w. and w/o using CBAM attention modules.

TPR for all participants exceed 95%, while for most participants, the FPRs are nearly 0%. The results demonstrate the effectiveness of *BioFace-3D* for user authentication.

Performance Under Blind Attack. While under blind attacks, the authentication system can successfully achieve nearly one hundred percent of rejecting illegitimate users. The findings are intuitive because it's extremely unlikely for attackers to correctly guess the order of the user's facial gestures. Even the attacker correctly guessed the series of gestures, it's nearly impossible to generate a similar biometric signal as the legitimate user. Therefore, TPR on the ROC curve is almost 100% and FPR is nearly 0%.

Performance Under Credential-aware Attack.

As illustrated in Fig. 12, under credential-aware attacks, our authentication system can also achieve an exceptionally low attack success rate, approaching zero. Even if the attacker possesses knowledge of the legitimate user's credentials, replicating the biosignals originating from the legitimate user's facial expressions is nearly impossible.

Impact of CBAM. We further evaluate the impact of CBAM on system performance. In Fig. 12 (a), the ROC curve depicts the model's performance when trained without CBAM layers under the credential-aware attack, with the anger gesture set as the credential. The corresponding AUC values for one gesture, three in a series, and five in a series are 0.9264, 0.9911, and 0.9956, respectively. Fig. 12 (b) illustrates the ROC curve of the model trained with CBAM layers, and we find out the results are increased to 0.9341, 0.9851, and 0.9995, respectively. This observation underscores the effectiveness of CBAM in enhancing the model's performance. Furthermore, we also check how adding CBAM affects facial landmark tracking. Specifically, we incorporate CBAM layers subsequent to each convolutional layer in the biosignal-based CNN network. We find that using CBAM leads to changes in the average MAE and RE. With CBAM, the average MAE and RE become 1.11% and 2.82%, which is an increase of 0.74% and 0.53%, respectively.

Impact of Series of Gestures. Fig. 13 shows the ROC curves of single gestures, and we found out that the fear gesture exhibits the best performance with an AUC of 0.9733. Fig. 14 presents the results from multiple trials, where users make repeated attempts to pass the system. Specifically, under the second trial, the average TPR could reach 0.99, and get close to 1 under the third trial. These promising results demonstrate the capability of *BioFace-3D* for user authentication.

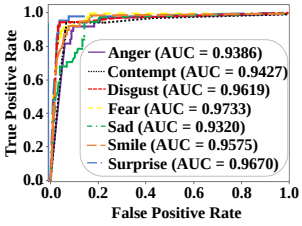


Fig. 13. ROC curves of user authentication with individual gestures.

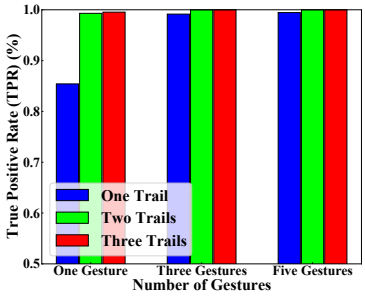


Fig. 14. True Positive Rate for multiple Trails with different number of gestures for all possible combinations.

7.3 Performance on User Identification

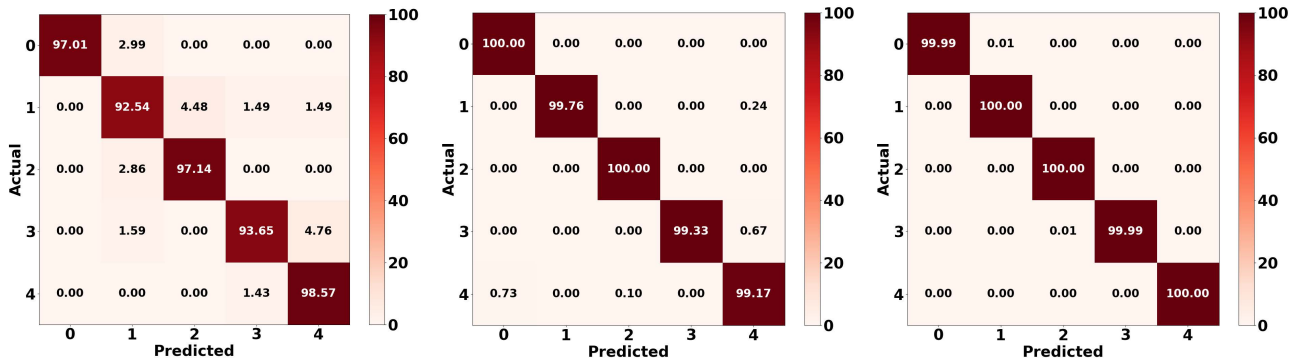
Fig. 15 illustrates the confusion matrices for use identification. As the series of gestures and soft majority vote method is deployed, the accuracy of identifying each user is raised significantly. As illustrated in Fig. 15 (a), when only one gesture is performed the overall accuracy is 93.89%. However, as shown in Fig. 15 (b) and (c), three gestures in series and five gestures in series can significantly increase the accuracy to 99.65% and 99.80%, respectively. This demonstrates that the robustness of *BioFace-3D* for user identification.

7.4 Impact of Data Augmentation.

We also evaluate the impact of data augmentation on user identification. Without data augmentation, the average accuracy for single, three, and five gestures is 66.25%, 89.96%, and 94.85%, respectively, which represent a reduction of 27.64%, 9.69%, and 4.95% compared to the results obtained with data augmentation. The results underscore the efficacy of employing data augmentation techniques.

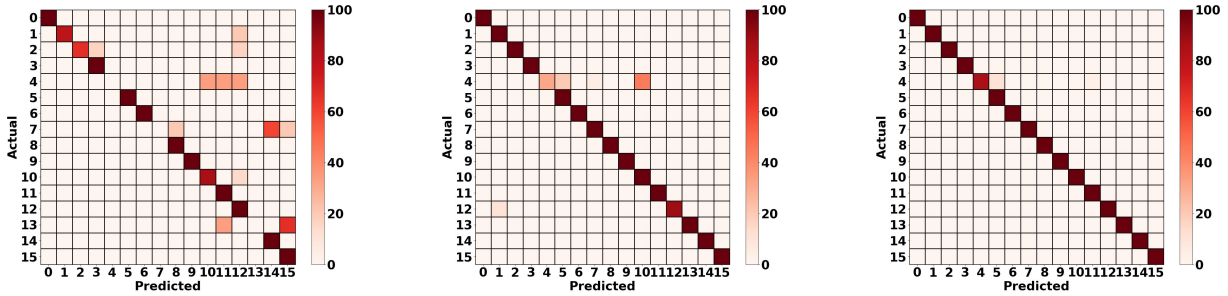
7.5 Performance of 16-User Identification.

We further evaluate the performance of 16-user identification. As illustrated in Fig. 16, the overall identification accuracy is 81.25%, 98.13%, and 99.87% with one, three, and five gestures in a series. The promising results demonstrate the effectiveness of user identification even when as many as 16 participants are involved in the system.



(a) Single-gesture based (b) Three-gesture based (c) Five-gesture based

Fig. 15. Confusion matrix of user identification of *BioFace-3D* with all possible gesture combinations.



(a) Single-gesture based (b) Three-gesture based (c) Five-gesture based

Fig. 16. Confusion matrix of 16-user identification of *BioFace-3D* with all possible gesture combinations.

8 Related Work

Camera-based Facial Landmark Detection. Traditional holistic camera-based methods [6], [7] detect facial landmarks by iteratively mapping a statistical facial model to the video frames. Constrained Local Model (CLM)-based methods [48], [49] build independent local shape models for each landmark, making them more robust to illumination and occlusion. Differently, deep-learning-based methods [8], [29], [50] extract high-level features from images and further learn a mapping to landmark locations via deep learning. However, these solutions require users to face a camera at all times without occlusions and under good lighting conditions.

Speech-driven Facial Animation. Early works [51], [52] utilize hidden Markov model (HMM) to generate speech-driven facial animations. Recent studies show a success of generating facial animations from audio spectrograms using 2D CNN [53] and from raw audio waveform using 1D CNN [54]. Moreover, LSTM-based methods have also been deployed for synthesizing mouth animations [55] or reconstructing full facial landmark [56]. However, these studies are not able to recognize silent facial gestures or expressions while talking.

Wearable-sensor-based Facial Movement Classification. Some studies recognize the user's facial movements using wearable sensors. For instance, speech-related movements can be sensed using capacitive sensors [10] or magnetic sensors attached to the tongue surface [9], [57], facial expressions can be identified using smart glasses with piezoelectric sensors [58] or optical sensors [59], and facial gestures can be sensed using earphone microphone [60], or acoustic interferometry [61]. Additionally, EMG & EEG signals have been shown effective in distinguishing a limited set of pre-defined facial gestures. Through attaching sensors around the user's eyes and forehead, previous studies can perform 5-class [62], 9-class [63], 10-class [64], 11-class gesture recognition [65]. More recently, Matthies *et al.* use tiny biosensors placed inside the ear canal to distinguish a set of 5 facial gestures [11], and Nguyen *et al.* use EMG signals captured behind the user's ears to sense tongue movements [66]. In addition to EMG, a few studies (e.g., [12], [21]) propose to use EOG signals to track eye movements to interact with machines. However, all these studies are classification-based methods and cannot be used for continuous 3D facial reconstruction.

Wearable/VR Authentication. EEG signals have been proved reliable on authentication users using commercial headsets(e.g., Emotiv Epoch+ [13]), leveraging machine-learning-based algorithms [14], [16] or least square estimation [15]. Alternatively, Google Glass could also serve as an authentication tool by harnessing its touchpad functionality [67] or internal camera [68]. Yang et al. introduce MotionAuth, an authentication scheme that leverages biometric data collected from a smart wristband [69], and Gafurov et al. incorporate IMU sensors into shoes to authenticate users by analyzing foot movements [70]. Regarding VR authentication, diverse biometric modalities have been employed including head motion [17], [18], [71], body motion [19], gaze movements [72], and ultrasonic reverberations caused by the shape of the users' heads [20]. However, none of these studies have the capacity to simultaneously perform 3D facial reconstruction while their sensor placement is quite obtrusive.

9 Conclusion

In this paper, we propose *BioFace-3D*, the first single-earpiece lightweight biosensing system for continuous 2D facial landmarks tracking, 3D facial animation rendering, and user authentication/identification. *BioFace-3D* can also generate personalized photo-realistic animations that incorporate specific facial details of the user. We design a novel cross-modal transfer learning framework to leverage high-precision camera sensor to guide the training of the biosensing model. We conducted extensive experiments involving 16 participants under various settings. The results demonstrated that the proposed *BioFace-3D* can accurately track major facial landmarks in a continuous manner with only 1.85 mm average error and 3.38% normalized mean error. Moreover, *BioFace-3D* can authenticate users with high accuracy, low false positive rate, and is robust to various types of attacks.

Acknowledgments

We would like to thank our anonymous reviewers for their insightful feedback. This work was supported in part by NSF grants ECCS-2132106 and ECCS-2132112.

References

[1] S. Carrasco and M. Á. S. UAH, "D3. 3 driver monitoring concept report," 2020.

[2] J. Kumari, R. Rajesh, and K. Pooja, "Facial expression recognition: A survey," Procedia Computer Science, vol. 58, pp. 486–491, 2015.

[3] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," Speech Communication, vol. 52, no. 4, pp. 270–287, 2010.

[4] Y. Wu, C. Shi, T. Zhang, P. Walker, J. Liu, N. Saxena, and Y. Chen, "Privacy leakage via unrestricted motion-position sensors in the age of virtual reality: A study of snooping typed input on virtual keyboards," in 2023 IEEE Symposium on Security and Privacy (SP), pp. 3382–3398, IEEE Computer Society, 2023.

[5] Ü. Meteriz-Yıldıran, N. F. Yıldıran, A. Awad, and D. Mohaisen, "A keylogging inference attack on air-tapping keyboards in virtual environments," in 2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp. 765–774, IEEE, 2022.

[6] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," Computer vision and image understanding, vol. 61, no. 1, pp. 38–59, 1995.

[7] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," IEEE Transactions on pattern analysis and machine intelligence, vol. 23, no. 6, pp. 681–685, 2001.

[8] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al., "Deep high-resolution representation learning for visual recognition," IEEE transactions on pattern analysis and machine intelligence, 2020.

[9] L. A. Cheah, J. M. Gilbert, J. A. González, P. D. Green, S. R. Ell, R. K. Moore, and E. Holdsworth, "A wearable silent speech interface based on magnetic sensors with motion-artefact removal.," in BIODEVICES, pp. 56–62, 2018.

[10] R. Li, J. Wu, and T. Starner, "Tongueboard: An oral interface for subtle input," in Proceedings of the 10th Augmented Human International Conference 2019, pp. 1–9, 2019.

[11] D. J. Matthies, B. A. Strecker, and B. Urban, "Earfieldsensing: A novel in-ear electric field sensing to enrich wearable gesture input through facial expressions," in Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 1911–1922, 2017.

[12] Y. Nam, B. Koo, A. Cichocki, and S. Choi, "Gom-face: Gkp, eog, and emg-based multimodal interface with application to humanoid robot control," IEEE Transactions on Biomedical Engineering, vol. 61, no. 2, pp. 453–462, 2013.

[13] Emotiv, "Emotiv epoch+." https://www.emotiv.com/epoc/, 2023.

[14] I. Jayarathne, M. Cohen, and S. Amarakeerthi, "Brainid: Development of an eeg-based biometric authentication system," in 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 1–6, IEEE, 2016.

[15] Q. Wu, Y. Zeng, C. Zhang, L. Tong, and B. Yan, "An eeg-based person authentication system with open-set capability combining eye blinking signals," Sensors, vol. 18, no. 2, p. 335, 2018.

[16] T. Koike-Akino, R. Mahajan, T. K. Marks, Y. Wang, S. Watanabe, O. Tuzel, and P. Orlik, "High-accuracy user identification using eeg biometrics," in 2016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC), pp. 854–858, IEEE, 2016.

[17] T. Mustafa, R. Matovu, A. Serwadda, and N. Muirhead, "Unsure how to authenticate on your vr headset? come on, use your head!," in Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics, pp. 23–30, 2018.

[18] M. Sivasamy, V. Sastry, and N. Gopalan, "Vrcauth: continuous authentication of users in virtual reality environment using head-movement," in 2020 5th International Conference on Communication and Electronics Systems (ICCES), pp. 518–523, IEEE, 2020.

[19] K. Pfeuffer, M. J. Geiger, S. Prange, L. Mecke, D. Buschek, and F. Alt, "Behavioural biometrics in vr: Identifying people from body motion and relations in virtual reality," in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–12, 2019.

[20] R. Wang, L. Huang, and C. Wang, "Low-effort vr headset user authentication using head-reverberated sounds with replay resistance," in 2023 IEEE Symposium on Security and Privacy (SP), pp. 3450–3465, IEEE Computer Society, 2023.

[21] C. S. L. Tsui, P. Jia, J. Q. Gan, H. Hu, and K. Yuan, "Emg-based hands-free wheelchair control with eog attention shift detection," in 2007 IEEE International Conference on Robotics and Biomimetics (ROBIO), pp. 1266–1271, IEEE, 2007.

[22] Y. Wu, V. Kakaraparthi, Z. Li, T. Pham, J. Liu, and P. Nguyen, "Bioface-3d: continuous 3d facial reconstruction through lightweight single-ear biosensors," in Proceedings of the 27th Annual International Conference on Mobile Computing and Networking, pp. 350–363, 2021.

[23] M. E. Tagluk, N. Sezgin, and M. Akin, "Estimation of sleep stages by an artificial neural network employing eeg, emg and eog," Journal of medical systems, vol. 34, no. 4, pp. 717–725, 2010.

[24] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2235–2245, 2018.

[25] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4d scans.," ACM Trans. Graph., vol. 36, no. 6, pp. 194–1, 2017.

[26] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1145–1153, 2017.

[27] P. Perkins and S. Heber, "Identification of ribosome pause sites using a z-score based peak detection algorithm," in 2018 IEEE 8th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS), pp. 1–6, IEEE, 2018.

[28] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001, vol. 1, pp. I–I, IEEE, 2001.

[29] Y. Wu, T. Hassner, K. Kim, G. Medioni, and P. Natarajan, "Facial landmark detection with tweaked convolutional neural networks," IEEE transactions on pattern analysis and machine intelligence, vol. 40, no. 12, pp. 3067–3074, 2017.

[30] X. Peng, R. S. Feris, X. Wang, and D. N. Metaxas, "A recurrent encoder-decoder network for sequential face alignment," in European conference on computer vision, pp. 38–56, Springer, 2016.

[31] W. Wu, C. Qian, S. Yang, Q. Wang, Y. Cai, and Q. Zhou, "Look at boundary: A boundary-aware face alignment algorithm," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2129–2138, 2018.

[32] A. T. Fairbanks and E. F. Fairbanks, Human proportions for artists. Fairbanks Art and Books, 2005.

[33] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1d convolutional neural networks and applications: A survey," Mechanical Systems and Signal Processing, vol. 151, p. 107398, 2021.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[35] U. Prabhu, K. Seshadri, and M. Savvides, "Automatic facial landmark tracking in video sequences using kalman filter assisted active shape models," in European Conference on Computer Vision, pp. 86–99, Springer, 2010.

[36] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 157–164, 2023.

[37] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp. 0–0, 2019.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016.

[39] Y. Ren, G. Li, Y. Chen, T. H. Li, and S. Liu, "Pirenderer: Controllable portrait image generation via semantic neural rendering," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13759–13768, 2021.

[40] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," J. Artif. Int. Res., vol. 16, p. 321–357, jun 2002.

[41] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in Proceedings of the European conference on computer vision (ECCV), pp. 3–19, 2018.

[42] A. Graves and A. Graves, "Long short-term memory," Supervised sequence labelling with recurrent neural networks, pp. 37–45, 2012.

[43] P. Ekman and D. Keltner, "Universal facial expressions of emotion," Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture, pp. 27–46, 1997.

[44] S. Zhu, C. Li, C. Change Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4998–5006, 2015.

[45] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 532–539, 2013.

[46] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 397–403, 2013.

[47] "Demo video for bioface-3d." https://mosis.eecs.utk.edu/bioface-3d.html, 2023.

[48] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models.," in Bmvc, vol. 1, p. 3, Citeseer, 2006.

[49] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," International journal of computer vision, vol. 91, no. 2, pp. 200–215, 2011.

[50] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in European conference on computer vision, pp. 94–108, Springer, 2014.

[51] M. Brand, "Voice puppetry," in Proceedings of the 26th annual conference on Computer graphics and interactive techniques, pp. 21–28, 1999.

[52] K. Choi, Y. Luo, and J.-N. Hwang, "Hidden markov model inversion for audio-to-visual conversion in an mpeg-4 facial animation system," Journal of VLSI signal processing systems for signal, image and video technology, vol. 29, no. 1, pp. 51–61, 2001.

[53] H. X. Pham, Y. Wang, and V. Pavlovic, "End-to-end learning for 3d facial animation from raw waveforms of speech," arXiv preprint arXiv:1710.00920, 2017.

[54] S. E. Eskimez, R. K. Maddox, C. Xu, and Z. Duan, "Noise-resilient training method for face landmark generation from speech," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 27–38, 2019.

[55] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," ACM Transactions on Graphics (ToG), vol. 36, no. 4, pp. 1–13, 2017.

[56] S. E. Eskimez, R. K. Maddox, C. Xu, and Z. Duan, "Generating talking face landmarks from speech," in International Conference on Latent Variable Analysis and Signal Separation, pp. 372–381, Springer, 2018.

[57] H. Sahni, A. Bedri, G. Reyes, P. Thukral, Z. Guo, T. Starner, and M. Ghovanloo, "The tongue and ear interface: a wearable system for silent speech recognition," in Proceedings of the 2014 ACM International Symposium on Wearable Computers, pp. 47–54, 2014.

[58] J. Scheirer, R. Fernandez, and R. W. Picard, "Expression glasses: a wearable device for facial expression recognition," in CHI'99 Extended Abstracts on Human Factors in Computing Systems, pp. 262–263, 1999.

[59] K. Masai, K. Kunze, D. Sakamoto, Y. Sugiura, and M. Sugimoto, "Face commands-user-defined facial gestures for smart glasses," in 2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pp. 374–386, IEEE, 2020.

[60] T. Amesaka, H. Watanabe, and M. Sugimoto, "Facial expression recognition using ear canal transfer function," in Proceedings of the 23rd International Symposium on Wearable Computers, pp. 1–9, 2019.

[61] Y. Iravantchi, Y. Zhang, E. Bernitsas, M. Goel, and C. Harrison, "Interferi: Gesture sensing using on-body acoustic interferometry," in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–13, 2019.

[62] M. Hamedi, I. M. Rezazadeh, and M. Firoozabadi, "Facial gesture recognition using two-channel bio-sensors configuration and fuzzy classifier: A pilot study," in International Conference on Electrical, Control and Computer Engineering 2011 (InECCE), pp. 338–343, IEEE, 2011.

[63] I. M. Rezazadeh, S. M. Firoozabadi, H. Hu, and S. M. R. H. Golpayegani, "A novel human–machine interface based on recognition of multi-channel facial bioelectric signals," Australasian physical & engineering sciences in medicine, vol. 34, no. 4, pp. 497–513, 2011.

[64] M. Hamedi, S.-H. Salleh, M. Astaraki, and A. M. Noor, "Emg-based facial gesture recognition through versatile elliptic basis function neural network," Biomedical engineering online, vol. 12, no. 1, p. 73, 2013.

[65] M. Hamedi, S.-H. Salleh, T. Tan, K. Ismail, J. Ali, C. Dee-Uam, C. Pavaganun, and P. Yupapin, "Human facial neural activities and gesture recognition for machine-interfacing applications," International Journal of Nanomedicine, vol. 6, p. 3461, 2011.

[66] P. Nguyen, N. Bui, A. Nguyen, H. Truong, A. Suresh, M. Whitlock, D. Pham, T. Dinh, and T. Vu, "Tyth-typing on your teeth: Tongue-teeth localization for human-computer interface," in Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services, pp. 269–282, 2018.

[67] J. Chauhan, H. J. Asghar, M. A. Kaafar, and A. Mahanti, "Gesture-based continuous authentication for wearable devices: the google glass case," arXiv preprint arXiv:1412.2855, 2014.

[68] R. Khan, R. Hasan, and J. Xu, "Sepia: Secure-pin-authentication-as-a-service for atm using mobile and wearable devices," in 2015 3rd IEEE International Conference on Mobile Cloud Computing, Services, and Engineering, pp. 41–50, IEEE, 2015.

[69] J. Yang, Y. Li, and M. Xie, "Motionauth: Motion-based authentication for wrist worn smart devices," in 2015 IEEE International conference on pervasive computing and communication workshops (PerCom Workshops), pp. 550–555, IEEE, 2015.

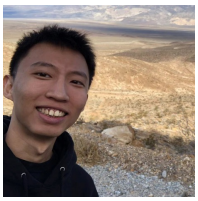
[70] D. Gafurov, P. Bours, and E. Snekkenes, "User authentication based on foot motion," Signal, Image and Video Processing, vol. 5, pp. 457–467, 2011.

[71] R. Miller, A. Ajit, N. K. Banerjee, and S. Banerjee, "Realtime behavior-based continual authentication of users in virtual reality environments," in 2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR), pp. 253–2531, IEEE, 2019.

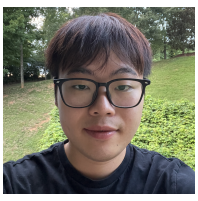
[72] J. Liebers and S. Schneegass, "Gaze-based authentication in virtual reality," in ACM Symposium on Eye Tracking Research and Applications, pp. 1–2, 2020.

[73] R. Ekman, What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA, 1997.

[74] U. Zarins, Anatomy of Facial Expressions. Exonicus, Incorporated, 2018.
[75] A. Banerjee, S. Datta, M. Pal, A. Konar, D. Tibarewala, and R. Janarthanan, "Classifying electrooculogram to detect directional eye movements," Procedia Technology, vol. 10, pp. 67–75, 2013.
[76] L. Learning, "Muscle contraction and locomotion." https://courses.lumenlearning.com/ivytech-bio1-1/chapter/muscle-contraction-and-locomotion/, 2021.
[77] W. Manjula, M. Sukumar, S. Kishorekumar, K. Gnanashanmugam, and K. Mahalakshmi, "Smile: A review," Journal of pharmacy & bioallied sciences, vol. 7, no. Suppl 1, p. S271, 2015.
[78] D. wearable sensors for movement sciences, "How to improve emg signal quality." https://delsys.com/emgworks/signal-quality-monitor/improve/, 2021.
[79] S. Karamizadeh, S. M. Abdullah, A. A. Manaf, M. Zamani, and A. Hooman, "An overview of principal component analysis," Journal of Signal and Information Processing, vol. 4, no. 3B, p. 173, 2013.
[80] T. Instruments, "Ads1299-x low-noise, 4-, 6-, 8-channel, 24-bit, analog-to-digital converter for eeg and biopotential measurements." https://www.ti.com/lit/ds/symlink/ads1299.pdf?ts=1615154540121, 2020.
[81] O. BCI, "Cyton biosensing board (8-channels)." https://shop.openbci.com/products/cyton-biosensing-board-8-channel?variant=38958638542, 2021.
[82] "Covidien kendall disposable surface emg/ecg/ekg electrodes 1" (24mm)." https://bio-medical.com/covidien-kendall-disposable-surface-emg-ecg-ekg-electrodes\-1-24mm-50pkg.html, 2021.
[83] "Monsoon high voltage power monitor." https://www.msoon.com/high-voltage-power-monitor, 2021.
[84] T. Gong, Y. Kim, J. Shin, and S.-J. Lee, "Metasense: few-shot adaptation to untrained conditions in deep mobile sensing," in Proceedings of the 17th Conference on Embedded Networked Sensor Systems, pp. 110–123, 2019.
[85] S. Park, S. D. Mello, P. Molchanov, U. Iqbal, O. Hilliges, and J. Kautz, "Few-shot adaptive gaze estimation," in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9368–9377, 2019.
[86] J. G. Webster, "Reducing motion artifacts and interference in biopotential recording," IEEE transactions on biomedical engineering, no. 12, pp. 823–826, 1984.



Yi Wu is a Ph.D. student in the Department of Electrical Engineering and Computer Science at the University of Tennessee, Knoxville. He received his B.E. and M.S. degrees from the University of Electronic Science and Technology of China and Rutgers University, respectively. His research interests include mobile sensing and cybersecurity.



Xiande Zhang is a graduate student pursuing his Master's degree in Computer Engineering at the University of Tennessee, Knoxville (UTK). His research interests are primarily focused on mobile sensing, system security, and the innovative application of smart devices. Prior to this, he earned his B.E. in Computer Engineering from the Department of Electrical Engineering and Computer Science.

Tianhao Wu is currently pursuing his Ph.D. degree at the Mobile Sensing and Intelligence Security (MoSIS) Lab, Department of Electrical Engineering and Computer Science at the University of Tennessee, Knoxville (UTK). He received his B.E. degree from the School of Electrical Engineering and Information, Northeast Agricultural University (NEAU). His current research interests include mobile sensing, human-computer interaction, and smart health.

Bing Zhou is a senior research engineer at Snap Research NYC. Before that, he was a research staff member at IBM T.J. Watson Research Center, Yorktown Heights, NY. His research interests are mobile sensing and computing, human computer interaction, 3D animation generation and location based services. He has published in top tier conferences as first or corresponding author such as MobiCom, MobiSys, SenSys, CHI, UbiComp, etc. Dr. Zhou obtained his PhD from ECE department, Stony Brook University in 2019. He received his Bachelor's degree in Applied Physics from University of Science and Technology of China (USTC) in 2011 and master degree from Chinese Academy of Sciences in 2014.

Phuc "VP" Nguyen is an Assistant Professor at Manning College of Information and Computer Sciences, University of Massachusetts Amherst. He is also affiliated with the Institute for Applied Life Sciences at UMass. He directs the Wireless and Sensor Systems Lab (WSSL). He is the recipient of SONY Faculty Innovation Award 2021, CACM Research Highlights 2020, 2021, ACM SIGMOBILE Research Highlights 2017, 2020, 2022, UTA CSE Pre-Tenure Research Award 2022, Best Paper Award at ACM MobiCom 2019, Best Paper Runner-up at ACM SenSys 2018, and Best Paper Nominee at ACM SenSys 2017. His research interests are mobile/wearable computing, wireless communication, and Internet of Things.

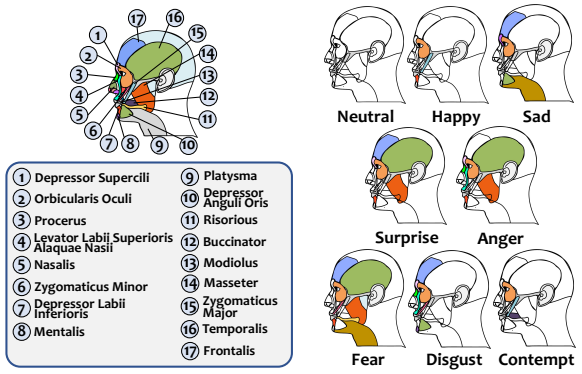
Jian Liu is an Assistant Professor in the Department of Electrical Engineering and Computer Science at the University of Tennessee, Knoxville (UTK). He leads Mobile Sensing and Intelligence Security (MoSIS) Lab @ UTK. His research interests span robust and trustworthy machine learning, computational sensing, system security, and smart healthcare. His research work has been published at top-tier security/mobile-computing/HCI/AI venues (e.g., ACM MobiCom, IEEE S&P, ACM CCS, CVPR, AAAI, ECCV, ACM MobiSys, ACM SenSys, ACM UbiComp, and ICASSP) and has been regularly featured in the media including BBC News, Yahoo News, MIT Technology Review, NBC New York, IEEE Spectrum, WCBS TV, and Voice of America TV, etc. He is the recipient of multiple awards, including two Best Paper Awards at IEEE SECON 2017 and IEEE CNS 2018, ACM SigMobile Research Highlights 2022, and ECE Graduate Program Academic Achievement Award at Rutgers, etc. He also filed seven U.S. patents, two of which have been licensed to industrial companies.

APPENDIX A
PRELIMINARIES

Facial Muscles and Eye Movements. Facial muscles, as illustrated in Fig. 17 (a), are striated skeletal muscles lying underneath the skin of the face and scalp to perform important functions for daily life, such as mastication and facial expressions. Different facial movements or expressions are produced by the contraction of a different set of facial muscles [73], [74]. For instance, *smile* involves a person pulling their lip corners up, thereby, raising their cheeks towards the eyes, making the eyelids come closer. These micro-facial movements are mainly driven by zygomaticus major, orbicularis oris, and orbicularis oculi. Differently, *surprise* involves raising eyebrows, widening eyes, opening the mouth, etc., which are usually associated with frontalis, depressor labii inferioris, temporalis, masseter, and orbicularis oris, etc. Fig. 17 (b) shows a common set of the activated facial muscles for seven universal expressions of emotion [74]. In addition, the eyeball acts as a dipole with a positive pole oriented anteriorly (cornea) and a negative pole oriented posteriorly (retina) [75]. This shows the potential of tracking the entire facial movements and eye movements through sensing the contraction of corresponding facial muscles and the bioelectrical signals caused by eye movements.

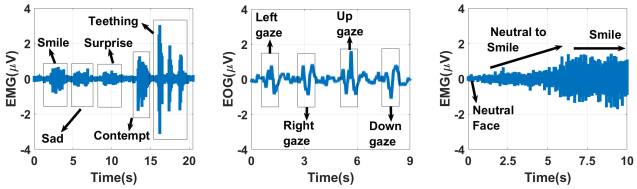
Sensing Facial Muscle Contractions via Single-ear Biosensors. Whenever a muscle contracts, a burst of electric impulses is generated which propagates through adjacent tissue and bone and can be recorded from neighboring skin areas [76]. These bursts of electricity can be captured by surface electrodes using electromyography (EMG) measurements if the electrodes are placed close to or on top of the activated muscles. Although the electrical potentials may pass through the connected muscles to be captured by an electrode, it remains unclear whether we can use the surface electrode attached to a least-obtrusive area, such as the area around one side of the ears, to sense the entire facial movements. We thus conduct an experiment where a surface electrode is attached to one side of the masseter around the ears while a participant performs multiple facial expressions including smile, sad, surprise, contempt, and chewing. In Fig. 18 (a), multiple events are generated corresponding to different muscle contractions. While some of them are not visually distinguishable due to the wide range frequency response of EMG, the events caused by facial activities can be clearly captured. We prove in Section 6 that the signals of each expression are indeed unique as validated by the Principal Component Analysis (PCA) presented. In the same setting, we ask the participant to look in different directions, and we observe that a unique voltage fluctuation is caused in the electrooculography (EOG) signals depending on the direction and duration of the movement, as shown in Fig. 18 (b). These observations confirm the possibility of using single-ear biosensors to sense the entire facial movements.

Continuously Sensing Muscle Contractions. To render continuous and smooth facial animations, the biosensors must be able to continuously track the muscle activities during the transitions between facial events. To validate the feasibility, we conduct an experiment to track the user



(a) Location of important facial muscles (b) Activated facial muscles during expressions

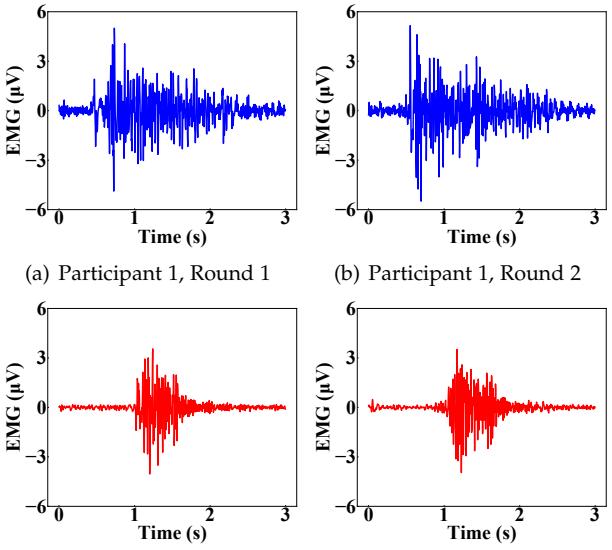
Fig. 17. Illustration of facial muscles.



(a) EMG signals of facial activities (b) EOG signals of eye movements (c) EMG signals of a slow smiling

Fig. 18. Biosignals collected from a side of masseter around the ears.

facial expression while the participant is asked to change their face from neural to smiling with a slower speed than normal (around 10 seconds). The purpose of this experiment is to validate whether the biosensor can capture muscle biosignals generated continuously during facial expression. Fig. 18 (c) shows the EMG signals obtained from the experiment, clearly validating the capability of surface electrode in continuously sensing facial activities.



(a) Participant 1, Round 1 (b) Participant 1, Round 2

(c) Participant 2, Round 1 (d) Participant 2, Round 2

Fig. 19. EMG signals corresponding to the facial gestures expressing "contempt", as performed by two users across multiple rounds.

APPENDIX B
FEASIBILITY STUDY FOR USER AUTHENTICATION & IDENTIFICATION.

The extracted EMG/EOG signals on the face during the same expression vary among individuals due to factors like distinct muscle activation patterns, variations in neuromuscular control, differences in facial anatomy, and the interplay of multiple muscles. These differences can also be influenced by physiological and psychological factors, leading to individualized patterns of muscle activation. For instance, two people smiling might show slightly different patterns of muscle activation in their cheek muscles (zygomaticus major) [77]. Hence, there exists significant potential in utilizing the biosignals extracted by *BioFace-3D* for the purpose of distinguishing and authenticating users. To validate this hypothesis, we conducted an experiment where two participants were required to perform the "contempt" gesture multiple times while wearing the sensing prototype. Fig. 19 (a) and (b), as well as (c) and (d), illustrate the EMG signals collected from participants 1 and 2, respectively. It is evident that when performing identical gestures, the EMG signals of different participants display entirely distinct patterns, whereas the biosignals of the same participant remain remarkably consistent. The observation demonstrates the feasibility of leveraging *BioFace-3D* for user authentication/identification.

APPENDIX C
SYSTEM IMPLEMENTATION

C.1 Electrode Placements

A traditional bio-electrical sensor channel includes three types of electrodes: *reference electrode*, *measurement electrode*, and *ground electrode*. To provide a relatively stable reference point and driven ground, the reference electrode and ground electrode should be attached to bony areas to keep all the underlying muscular signals minimized. Thus, we attach these two types of electrodes to the back of the ears (i.e., mastoid bone) in the design of *BioFace-3D*. Regarding the measurement electrode, from our analysis in terms of the unobtrusiveness and the capability of sensing, it could be placed at six locations P1-P6 as illustrated in Fig. 20. In particular, P1 is on the temporalis, proximity to orbicularis oculi; P2 is on the temporalis and temporal bone, proximity to deeper head; P3 is on the masseter (on the zygomatic bone); P4 is at the junction of the mandible and temporal bones with proximity to temporalis and masseter; P5 is at the junction of risorius, masseter, platysma, on the mandible bone; and P6 is on the lower side of the masseter, proximity to risorius and platysma. To find the most suitable location for the measurement electrode, we perform both SNR and Principal Component Analysis (PCA) analyses below.

SNR Analysis. We calculate the Signal-to-Noise Ratio (SNR) of the signals generated by each of the universal facial expressions using the six measurement electrode locations. To ensure the acceptable quality of the measured biosignals, the SNR should be greater than 1.2 db [78]. However, from our experiments we observe that a value more than 1.6 dB is acceptable to withstand the baseline noise variations. The results are shown in Table 2. We observe that P4-P6 have a relatively low SNR ($<$ 1.6 dB) for some of the expressions. For instance, *sad* has a low magnitude at P4 because the location is situated outside the masseter, which has loose connections to the depressor anguli oris that facilitates the facial gesture. *Fear* has low magnitude at P6 as it is at the lower end of the masseter that has no connection to the muscles deforming the mouth. Through this analysis, we found that P1, P2, P3 locations perform well with all the universal facial expressions.

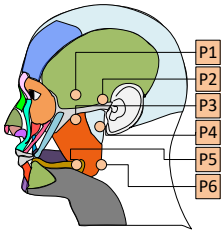


Fig. 20. Potential electrode placements.

TABLE 2
SNR results from the six facial locations in decibels (dB).

	Happy	Sad	Angry	Surprise	Fear	Disgust	Contempt
P1	12.89	4.24	7.00	4.87	7.07	1.98	16.50
P2	8.49	2.95	7.06	7.65	5.23	1.86	8.55
P3	10.18	5.50	7.00	4.87	7.07	1.98	16.50
P4	3.70	***1.58***	3.19	5.31	2.50	***1.27***	12.06
P5	6.58	11.97	4.19	3.14	2.60	***0.89***	16.24
P6	3.82	3.47	3.86	3.13	***1.54***	***0.56***	11.18

PCA Analysis. Although SNR is a great indicator for detecting facial activities, it does not provide sufficient details on the quality of the signals in distinguishing different facial activities. To analyze the distinguishability of the captured signals of different facial movements, we transform the gestural signals from P1, P2 and P3 locations to the frequency domain using Discrete Fourier Transform (DFT). The DFT signals are then projected into new dimensional space for feature engineering via Principal Component Analysis (PCA) separability scores [79]. The overall separability scores at P1, P2, P3 are 94.25%, 93.53%, 92.27%, respectively. This result affirms the fact that each facial movement generates a unique physiological signature at each of these facial locations. Specifically, P1 and P2 are affected by an overlap between *fear* and *anger* gestures while P3 is affected by an overlap between *smile* and *contempt* gestures. P3 can distinguish *fear* and *anger* due to its connections to frontal face muscles while P1 and P2 can separate *smile* and *contempt* as they can capture buccinator and zygomatic major activations in a fine grained manner. Due to the intrusive nature of P1, we choose to use two measurement electrodes at P2 and P3, which can complement to each other to sense the entire facial activities.

C.2 Prototype

Single Ear-piece Design. From our experiments, the gestural signals generated on both sides of the face are observed to be very similar in magnitude, shapes, etc. In particular, there are no significant changes to the dimensional space and separability scores of the gestural signals after PCA. For P2, the dimensional space has 270, 272, and 272 components when we use the data from left side, right side and both sides of the face, respectively. For P3, the dimensional

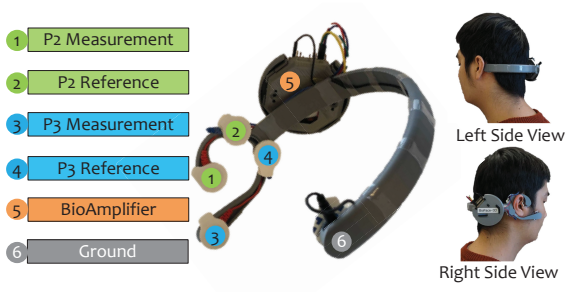


Fig. 21. *BioFace-3D* prototype.

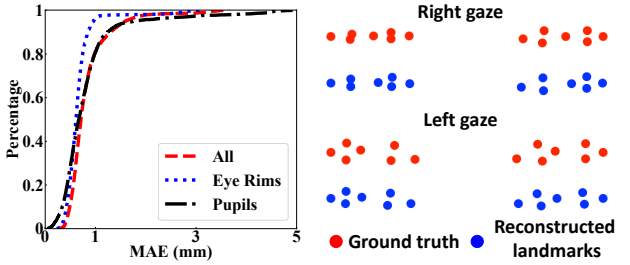
space has 153, 156, and 158 components. Hence, there are almost no unique features that can be added by the data from the second side of the face. The dimensional space explains 95% of the variance of the dataset and the separability scores for each case does not vary by more than 1% while remaining higher than 92%. Thus, universal gestures that involve muscle groups from both sides of the face can be captured with equal detail from electrode channels being placed on just one side of the face.

Prototype. The *BioFace-3D* wearable device is customized based on (a) dimension of the user's head (b) preference for the side of the earpiece. The earpiece design is dictated by the facial locations of measurement electrodes P2 and P3 as described previously. The reference electrodes are placed on a bony surface behind the ear such that those electrodes are sufficiently away from the facial muscle activity that the measurement electrodes capture. The earpiece provides slots for measurement, reference and ground electrode placements at precise locations as illustrated in Fig. 21. This earpiece is integrated with a headband that goes around the neck. We designed three sizes of prototypes that place the sensors in appropriate facial locations for three adult population groups: Large, Medium, and Small. For each of the sizes we designed two variants based on which side the earpiece is present. This allows for a wearable device that suits a large population. This headpiece also houses a circuit box to contain the hardware. All of the components in the headset are manufactured by 3D printing of PLA to ensure that the prototype is lightweight. *BioFace-3D* uses an ADS1299 based bio-amplifier circuit, i.e., OpenBCI [80], [81], and Ag/AgCl surface electrodes [82] that stick to the user's skin, as illustrated in Fig. 21. A Bluetooth module is integrated for data streaming. Due to the customized shape of the prototype, which is tailored for wearing around the user's ears, its design does not accommodate wearing on other body parts. We choose this part to maintain a minimized obtrusiveness level to the wearer. Sensing facial movements using sensors attached to other body parts (e.g., the chin and neck) is left as our future work.

APPENDIX D
PERFORMANCE ON FACIAL TRACKING

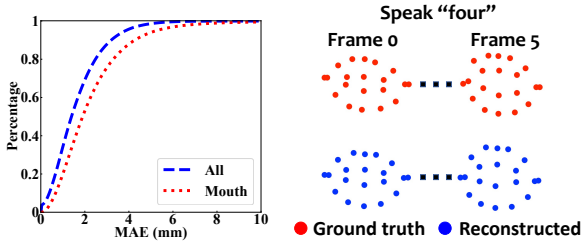
D.1 Eye Movement Tracking

To better evaluate the performance of gaze tracking, we collected another dataset involving three participants, who were asked to repeatedly look into four different directions (i.e., left, right, up, and down) for 300 seconds. Each



(a) MAE CDF for eye-related landmarks (b) Ground truth & reconstructed landmarks

Fig. 22. Performance of continuous eye-tracking.



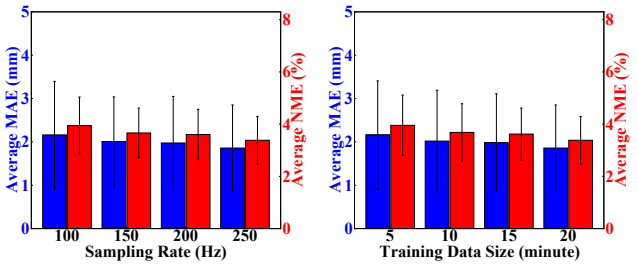
(a) MAE CDF for mouth-related landmarks (b) Ground truth & reconstructed landmarks

Fig. 23. Performance of continuous mouth movement tracking while the user is speaking.

gazing activity lasts for 2 seconds and was separated by 1 second *looking straight ahead*, which results in a total of 100 gaze movements. For each participant, we use the first 4 minutes for training and the remaining 1 for testing. The MAE CDF is shown in Fig. 22 (a), in which we achieve an average MAE of 0.82 mm for all eye-related landmarks, 0.73 mm for eye rims, and 0.95 mm for pupils. We found that 80% of the pupil landmarks have an error lower than 0.98 mm, which shows the promising capability of *BioFace-3D* for gaze tracking even in this active eye-moving setting. Examples of the reconstructed landmarks (right/left gaze) are shown in Fig. 22 (b).

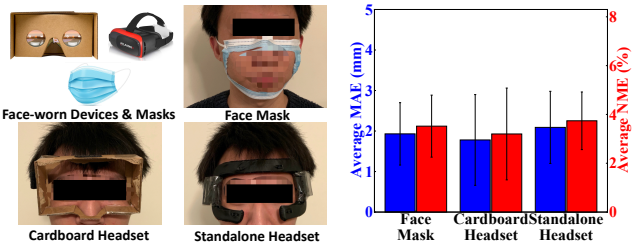
D.2 Facial Landmark Tracking (Speaking).

To comprehensively evaluate our system, we extended our experiments to other types of facial movements (i.e., speaking) by involving five participants who were asked to repeatedly speak nine digits (i.e., one to nine). During experiments, each digit was repeatedly spoken for 4 minutes, which results in a total of 36 minutes of data. We used 26 minutes of data for training and the remaining 10 minutes data for testing. The CDF curves for MAE are shown in Fig. 23 (a), in which we achieve an average MAE of 1.63 mm for all facial landmarks, while 2.39 mm for mouth-related landmarks. We found that 80% of the mouth landmarks have an error lower than 3.29 mm, which shows the promising capability of *BioFace-3D* in tracking facial movements of speaking. Examples of the reconstructed mouth landmarks at a interval of five frames of speaking *four* are shown in Fig. 23 (b). The promising results demonstrate the capability *BioFace-3D* of tracking the users' mouth movements while they are speaking, potentially extending our system to other usage scenarios such as speech enhancement.



(a) Impact of sampling rate (b) Impact of training data size

Fig. 24. Performance of facial landmark tracking with different sampling rate & training data size.



(a) Different facial occlusions (b) Landmark tracking error

Fig. 25. Performance of continuous facial landmark tracking under the presence of facial occlusions.

D.3 Micro-benchmark Tests

D.3.1 Impact of Biosensor Sampling Rate

To evaluate the impact of sampling rates on our system, we down-sample the frequency of the biosignal collected at 250 Hz to 50-200 Hz.

Fig. 24 (a) presents the average MAE and NME when varying the sampling rate from 100 Hz to 250 Hz. We observe that high sampling rate slightly improves the performance, and *BioFace-3D* is not very sensitive to changes in the sampling rate, given the range from 100 Hz to 250 Hz. Even if the sampling rate is decreased to 100 Hz, *BioFace-3D* still achieves an average MAE of 2.16 mm and NME of 3.94%, with average standard deviations of 1.23 mm and 1.09%, respectively. These results show that our system can also provide good performance even with a lower sampling rate, which can further reduce the computational complexity and power consumption.

D.3.2 Impact of Training Data Size

We then evaluate the system robustness with different training data sizes to seek the potential of further reducing training efforts. Fig. 24 (b) presents the overall system performance when varying the training data size from 5 minutes to 20 minutes for each participant, while all the remaining data is used for testing. We observe that even if the size of training data is decreased to 5 minutes, *BioFace-3D* still achieves an average MAE of 2.17 mm and NME of 3.95%. A larger training size would lead to better accuracy, but it remains operable if a user intends to have a quick enrollment process.

D.3.3 Impact of Face-worn Devices and Masks

Wearing face-worn devices/masks involve external forces (e.g., rubber bands for face coverings), which would tighten facial muscles and add additional pressure on the prototype, potentially introducing noises to biosensor readings. We further test the system performance with the presence of a face mask or a VR headset (a cardboard headset or a standalone headset), as shown in Fig. 25 (a). Specifically, we asked a participant to wear a face mask and two types of VR headset (i.e., a cardboard headset and a standalone headset) respectively while using our system. The training data is the 20 minutes data with no occlusion involved. To obtain the ground truth from the vision-based network, we cut off the front side of the mask to expose the mouth of the user, and tear off the headsets to reveal the user's eyes & eyebrows, as shown in Fig. 25 (a). Fig. 25 (b) presents the system performance when wearing face masks and head-worn VR headsets. Although wearing face-worn devices/masks decreases the performance, the overall performance remains within an acceptable range, e.g., average MAE of 1.93 mm, 1.78 mm, and 2.09 mm while wearing a face mask, a cardboard headset, and a standalone headset, respectively. These results demonstrate the robustness of *BioFace-3D* with different facial occlusions.

D.3.4 Resilience to Bursty Head Movements

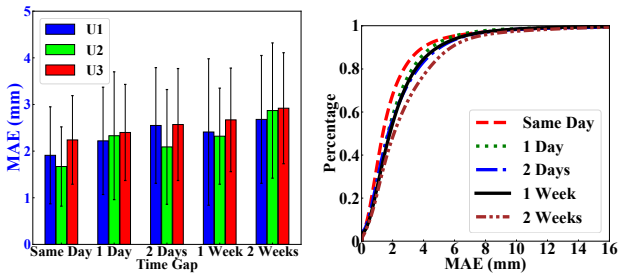
We are also interested in how bursty head movements impact our system. Specifically, the participant was asked to regularly rotate & shake his head during testing data collection, while ensuring the head could be captured by camera for the ground truth acquisition. The training data is the 20 minutes data with no head movements involved. With an average MAE of 1.79mm and an NME of 3.25%, BioFace-3D is resilient against active head movements, making it applicable to many practical scenarios involving active head movements.

D.3.5 Temporal Stability

The sensor measurement would be influenced by the day-by-day change of the users' body status, uncontrollable impurities on the skin surface, and the sensor displacement as the prototype won't be worn in exactly the same way. As time passes by, these issues may become more serious and therefore affect the sensor measurements at a greater scale. It is thus important to validate the system's temporal stability to prevent repetitive training. We asked three participants to collect another four sets of testing data (10 minutes each) which is separated from training data by 1 day, 2 days, 1 week, and 2 weeks. As shown in Fig. 26 (a), we found that in the worst case, *BioFace-3D* stills reaches an MAE of 2.87 mm over two weeks and there is no significant performance change in two-week period, as illustrated in Fig. 26 (b). These results affirm the fact that the sensitivity to sensor placement positions, which tend to differ minutely with each usage, have a negligible effect on the system outputs.

D.3.6 Computational Cost & Power Consumption

The inference time of 53 facial landmarks is measured on a single NVIDIA GTX 2080Ti GPU, and our model only takes around 0.033 ms to reconstruct a single frame, which is sufficient for real-time applications. Additionally, the model only takes around 0.775 ms to reconstruct the



(a) MAE for cross-day setting (b) CDF for cross-day setting

Fig. 26. Performance of landmark tracking over time.

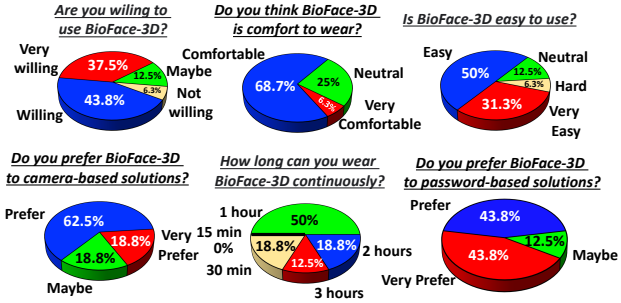


Fig. 27. Results of user study questionnaire.

3DMM of a single frame, and 0.038 ms for user authentication/identification. In addition, we use a power monitoring device (i.e., Monsoon High Voltage Power Monitor [83]) to measure the power consumption of *BioFace-3D*. All measurements are conducted at 60°F with a normal Lipo battery voltage (3.7V). Specifically, if the system is in the idle state where MCU is working in the idle mode without streaming data via Bluetooth, *BioFace-3D* consumes 118 mW on average. If *BioFace-3D* is sensing and streaming biosignal data via Bluetooth, the whole system's power consumption is 138 mW. This indicates that *BioFace-3D* can provide continuous data logging for 8.2 hours using a 500 mAh Lipo battery, which meets the requirements of most applications.

D.4 User Study

We asked the participants to fill a questionnaire, as shown in Fig. 27, on their experience with *BioFace-3D* after the experiments. We found that 81.3% of the participants are willing to use *BioFace-3D* and 75% of the participants feel it's comfortable to wear. 50% of the participants think *BioFace-3D* is easy to use and 31.3% of the participants feel it's very easy to use. We only got one negative feedback towards *BioFace-3D* , simply because the specific participant "doesn't want to have anything around his head". Additionally, 81.3% of the participants prefer *BioFace-3D* rather than traditional camera-based solutions, mostly due to the reason that *BioFace-3D* is more privacy preserving, can detect facial expression independently of body movements, and is reliable when parts of the face are blocked. All participants can use it for more than 30 minutes and 81.3% of the participants can use it more than 1 hour, which is sufficient for many usage scenarios. Specifically, the major reason which made 18.8% of the participants only choose to wear it for 30 minutes is the lack of adjustability. Due to the size of the prototype being fixed at the current stage, sometimes it cannot fit the user's head very appropriately and will cause displacement as time passes by, which may downgrade user experience. We plan to address this issue by utilizing more flexible materials to enhance the size variability. This is considered as our future work. Finally, 87.5% of the participants prefer to use *BioFace-3D* for authentication/identification compared to password-based solutions, due to its convenience and reliability.

APPENDIX E
DISCUSSION

User-independent Model. As the signal strength, response, and sensitivity of biosignals may vary from user to user, we currently adopt user-specific training to mitigate this variance and improve system accuracy. However, this might reduce the usability as new users have to undergo the enrollment phase before using the system. To improve usability, we can potentially train a generic user-independent model using data collected from a large set of users. When new users are introduced, the generic model can be adapted to the users with few calibration samples via meta-learning-based few-shot adaptation [84], [85]. We leave this as our future work.

Effects of Body Movements. Motion artifacts are a formidable noise that occurs in all Electrogram measurements [86]. It is a low-frequency noise occurring in the EOG frequency range. They occur due to two reasons: 1) relative motion between the surface electrodes and the skin surface; and 2) connection quality fluctuations between the wires and electrodes. We ensured that motion artifacts are mitigated by the design of the prototype which maintains the contact quality of the surface electrodes and keeps the connecting wires very short. This is evidenced by the evaluation of the system under rapid head movements in Appendix D.3.4. The body movements such as walking would have significantly less impact on the results as they introduce less relative motion between electrodes and skin as well as electrodes and wires when compared to rapid head movements. We note that the participants were allowed to move freely during our experiments as long as they can be captured in the video.

Potential Applications. Without requiring a camera positioned in front of the user, our system would introduce new opportunities in various emerging applications. For instance, through increasing the awareness of the user's real-time facial expressions and emotional states, our system can enable a more immersive user experience for existing AR/VR applications (e.g., face-to-face interactions), assess student engagement for online courses, and assist with driver fatigue detection to monitor abnormal behaviors, etc. In addition, our system can serve as a silent-speech interface for human-computer interaction. Through performing different facial gestures, people can interact with smart home appliances (e.g., turn down the volume of a smart speaker) and disabilities can control their handicap equipment (e.g., a wheelchair) more conveniently. Furthermore, our system can function as a continuous authentication mechanism for immersive virtual reality devices, such

as the Apple Vision Pro and Oculus Quest. This approach offers significant advantages over traditional password-based authentication methods, providing enhanced convenience and user experience. We plan to develop an API library which is compatible with major AR/VR platforms (e.g., OpenVR) and a mobile app to support various mobile devices in our future work.

Reducing Power Consumption. The current prototype can support up to 8.2 hours of continuous usage if paired up with a 500 mAh Li-ion battery. In our future work, we seek to further improve the system's energy efficiency by designing a customized data collection board using a more compact analog-to-digital converter with fewer channels for the biopotential measurements (e.g., ADS1299-4 consumes 43% less power than ADS1299-8 used in the current bio-amplifier circuit [80]).