IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 December 2022, accepted 4 February 2023, date of publication 8 February 2023, date of current version 23 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3243287

RESEARCH ARTICLE

A Selective Expression Manipulation With Parametric 3D Facial Model

JIWOO KANG[1,2], **HYEWON SONG**[3], **KYOUNGOH LEE**[4], **AND SANGHOON LEE**[3,5], **(Senior Member, IEEE)**

[1]Department of IT Engineering, Sookmyung Women's University, Yongsan-gu, Seoul 04310, Republic of Korea
[2]Design Research Institute for Creativity and Convergence, Sookmyung Women's University, Yongsan-gu, Seoul 04310, Republic of Korea
[3]Department of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, Republic of Korea
[4]AI Application Section, Electronics and Telecommunications Research Institute, Yuseong-gu, Daejeon 04129, Republic of Korea
[5]Department of Radiology, College of Medicine, Yonsei University, Seoul 03722, South Korea

Corresponding author: Sanghoon Lee (slee@yonsei.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (Ministry of Science and ICT (MSIT) (2020R1A2C3011697, 2022R1F1A1068704), the Yonsei University Research Fund of 2021 (2021-22-0001), and the Sookmyung Women's University Research Grants (No. 1-2203-2043).

ABSTRACT This paper proposes a novel method to represent expressive 3D facial shapes called the Selective Expression Manipulation (SEM) by fitting the expression coefficients of delta-blendshapes, which is the standard parametric facial model widely used in industries. SEM focuses on preserving blendshape semantics to characterize the facial shapes since the facial shape obtained by minimizing the distance to sparse facial landmarks might fail to signify a facial expression from a human being's perspective. Assuming each delta-blendshape corresponds to a facial movement with semantic meaning, SEM finds a series of facial motions required to compose the target facial expression. In addition, SEM sequentially determines a sufficient number of expressions and coefficients closely resembling the target facial movements by introducing similarities to quantify the directional correlation of facial motions between a target and a blendshape, excluding redundant expressions in terms of motions from the neutral shape. As a result, far fewer inter-correlated expressions that significantly increase the target correlation can be obtained. Furthermore, SEM exhibits substantial improvement in accuracy, correlation, semantics, and stability in experiments over previous facial fitting schemes and state-of-the-art methods. It is demonstrated that SEM enables accurate and realistic 3D facial shape generation by semantically manipulating expression delta-blendshapes.

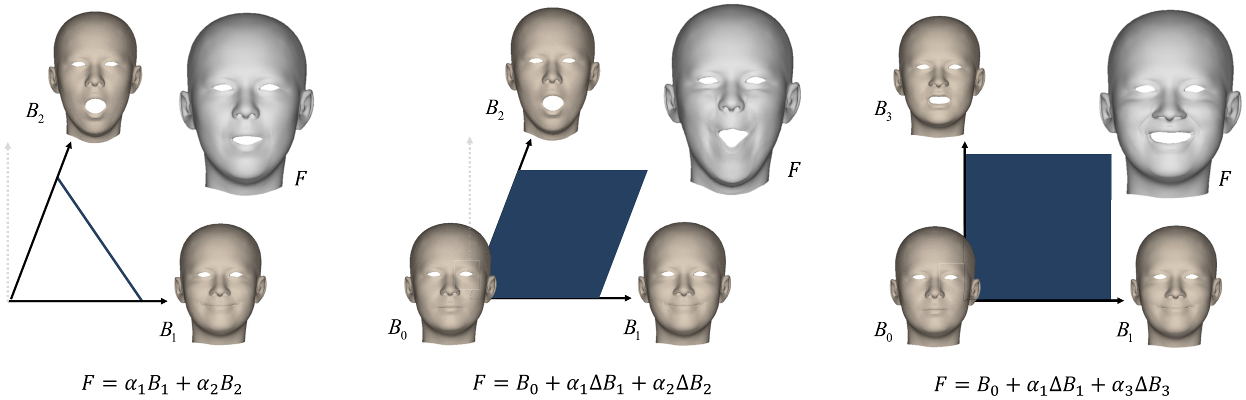
INDEX TERMS Human modeling, computational parameter fitting, 3D facial modeling, shape analysis, mesh models.

I. INTRODUCTION

3D human characters, in particular, 3D faces, are essential elements in virtual and augmented realities [1], [2]. Blendshape, a linear facial expression model, has been widely used in various applications to generate realistic 3D human faces such as face recognition [3], [4], face tracking [5], [6], [7], performance-based facial capture [8], [9], human reconstruction [10], [11], [12], [13], and facial retargeting [14], [15] thanks to its simplicity and expressiveness. In general, there are two types of blendshape representations: global-blanedshape and delta-blendshape. In the global-blendshape representation, as shown in Figure 1(a), a facial shape is generated in the span of the "whole"-facial blendshapes [16]. This approach ensures that the generated shape lies within a valid range of expressions, resulting in stable shapes and preventing unexpected expressions. Nevertheless, the concurrent representation of several expression models dilutes the expression of each model in the global-blendshape. To overcome this drawback, many researchers [17], [18], [19] used the "delta"-blendshape representation to overcome this drawback along with expression blendshapes that are constructed according to the Facial Action Coding System (FACS) [20]. In the delta-blendshape

The associate editor coordinating the review of this manuscript and approving it for publication was Claudia Raibulet.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/



$F = \alpha_1 B_1 + \alpha_2 B_2$

(a) Global-blendshape

$F = B_0 + \alpha_1 \Delta B_1 + \alpha_2 \Delta B_2$

(b) Delta-blendshape

$F = B_0 + \alpha_1 \Delta B_1 + \alpha_3 \Delta B_3$

(c) SEM

FIGURE 1. The approaches of blendshape representations. The sum of the delta-blendshape weights is one, i.e, $\sum \alpha_i = 1$. The delta-blendshape ΔB_i is defined as $B_i - B_0$. The blue line or squared region depicts the span of a facial shape generated by the representation. (a) Global blendshape, (b) Delta-blendshape, and (c) Selective Expression Manipulation (SEM).

representation, as shown in Figure 1(b), a facial shape is characterized in the span of the delta-shapes, which are the differences in the facial shape of each expression blendshape from a neutral expression blendshape. The delta-blendshape representation enables the facial shapes to be described with a wide variety of facial expressions. However, the combinations of these expression blendshapes do not always preserve the semantic meaning of the original expression, as depicted in Figure 1(b), because it uses many redundant expressions with high inter-correlations. When expression delta-blendshapes are being fitted to the point clouds or facial landmarks of a target, the expressions of the blendshapes may be locally overshot to make the facial shape closer to the fitting points. For these reasons, a combination of highly correlated blendshapes can cause an exaggerated and unstable facial shape. Thus, the performance degradation is inevitable.

In this paper, we propose a robust selective fitting approach for expression delta-blendshapes, called Selective Expression Manipulation (SEM). We assume that the expression blendshapes are not semantically duplicated with the others. In other words, each expression blendshape is semantically unique and it has a distinct expression. The key idea is that each expression delta-blendshape is considered as a facial motion which is the difference from the neutral expression blendshape. From this idea, we suppose that an expressive target face is formed by taking a series of facial motions from the expressionless or non-posed face (i.e., neutral expression face), as described in Figure 2. We can accumulate expression delta-blendshapes without losing facial expressiveness by composing facial motions that rarely correlate with each other. Thus, in the proposed method, a set of expression delta-blendshapes and their coefficients are obtained that can preserve the semantic meanings of their facial expressions by seeking the set that are least correlated to each other.

The *expression selection* and *exclusion* methods are proposed for SEM to highly correlate a combination of the expression delta-blendshapes with the target motion while decreasing the inter-correlations of the expression delta-blendshapes. The two measurements, *expressional* and *relative similarities*, are introduced to quantify the directional similarity of facial motions between the target and the expression delta-blendshape. Based on similarity measures, SEM obtains expressions that accurately represent the target's facial motions in the order of *expressional similarity*. The *relative similarity* allows SEM to compute the appropriate coefficients of the expression delta-blendshapes. SEM selects an expression with the highest similarity among the expression delta-blendshape candidates to update the facial shape in a greedy manner [22], [23]. Meanwhile, SEM excludes redundant expressions from the remaining candidates to represent the target face with less inter-correlated expressions by ensuring the selected expressions as a series of motions from the initial shape in a backtracking manner [24], [25]. We prove that the most motionally correlated to the target can be obtained by our greedy formulation that iteratively synthesizes a facial shape with the selected expression delta-blendshapes. Through exhaustive experiments to evaluate the accuracy, redundancy among expression delta-blendshapes, uniqueness, and semantics, including comparisons with state-of-the-art methods and baseline methods that cover the previous facial fitting schemes, it is demonstrated that SEM finds appropriate expressions to uniquely compose the expressive target shape by manipulating expression delta-blendshapes semantically, enabling accurate and realistic facial shape generation. In summary, we propose a selective expression delta-blendshape manipulation scheme, in which

- *Expressional and relative similarities* are introduced on the target shape and the expression delta-blendshapes to measure motional correlations under the assumption that the facial shape is generated by the motions from the neutral shape,
- *Expression selection* of most similar expressions is carried out in a greedy manner to address the expression

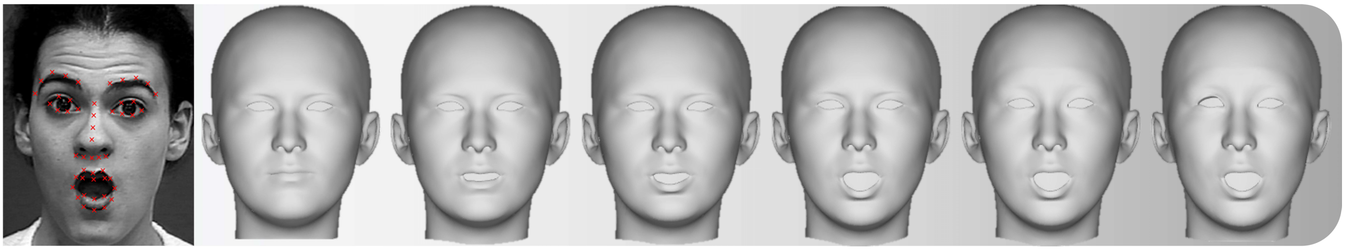


FIGURE 2. The proposed method selectively picks delta-blendshapes as a series of facial motions to semantically reveal a target expression under the assumption that each delta-blendshape is a facial movement needed to compose the target expression. Fitting the blendshapes for the given target points (marked by "x" on the face) by stacking the less inter-correlated expressions facilitates the representation of the facial shape semantically and expressively. The image was obtained from the CK+ database (©Jeffrey Cohn) [21].

fitting as finding the highest correlated motion to the target, and

- *Expression exclusion* of redundant expressions is performed in a single-step backtracking manner to represent the expressive target face with a combination of less inter-correlated expressions, enabling SEM to generate realistic facial shapes with significantly increased semantics and uniqueness.

II. PREVIOUS WORKS

Many other efforts have been made to overcome the drawbacks of the delta-blendshape approaches. In studies by several authors [26], [27], [28], [29], the $\mathbb{L}_1$ norm of blendshape weights have been used to achieve sparse weight activations, resulting in a smaller number of expressions. Also, the $\mathbb{L}_2$ regularization has been used in several works [30], [31], [32], [33], [34] to prevent the over-fitting of the expression models to the target face. Although these norm-based regularizations are useful techniques to prevent the expression of the facial shape from being exaggerated, they can cause the facial shape less expressive than the target face. This trade-off between shape stability and expressiveness is highly data-dependent, so the performance might be very sensitive to the individual and its expression, even when the same regularization definition is used. Simply increasing the sparsity of the expression coefficients (e.g., by penalizing the number of the expressions used to fit without considering the correlations between the expressions) does not help the facial shape become expressive and interpretable. Moreover, norm-based regularizations do not flexibly handle the semantic meaning of expression models.

Some methods used different types of prior constraints to regularize the blendshape expressions [35], [36], [37], [38], [39]. While these approaches can perform well, their quality depends heavily on the priors, which are selected manually, or requires a sufficient number of training examples. Nevertheless, these types of regularization have not been guaranteed as a fundamental solution to overcome the drawbacks of the delta-blendshape, as they still use whole-expression models, which are highly correlated with each other. Indeed, we found that using a selected set of expressions that are rarely correlated with each other can describe a target's face more expressively, while preserving uniqueness and interpretability by maintaining the semantics of each expression blendshape.

Parametric facial models, such as the 3D Morphable Model (3DMM) [30], [40], represent a facial shape on a low-dimensional face subspace by using the Principal Component Analysis (PCA). In recent works [33], [41], [42], [43], [44], [45], [46], [47], [48], facial geometry and reflectance have been estimated using the parametric facial models, but the baseline is on a neutral target face and ignores the target expression. In other works [49], [50], [51], [52], [53], [54], [55], an additional basis was used to model expressions parametrically, and in [30], both facial appearance and expression variations of in-the-wild facial shapes were modeled concurrently on a basis model. Although decomposition methods such as PCA, Independent Component Analysis (ICA), and Linear Discriminant Analysis (LDA), help obtain orthogonal basis vectors, a dimension transfer makes the basis vectors of the facial shape difficult to interpret and apply [16]. The decomposition alters the basis vectors, which have been commonly modeled on the basis of the facial unit actions [20], from local to global deformations. Although decomposition enables to model the deformation globally by capturing correlations across faces in the database used for training the model, global deformations depend on the statistics of the training set. They tend to require a sufficient number of coefficients to represent high-frequency details of the facial shape. Some works have shown that the data variance of facial movement is not jointly Gaussian and thus PCA is not sufficient to model expression variation [16], [56], [57]. Statistical priors should be required to fit the target more stably [58], [59]. Therefore, expression blendshapes without decomposition have been used to manipulate the blendshape expression coefficients in many related works [8], [9], [27], [28], [36], [60], [61], [62], [63].

Recently, Kang et al. used a facial fitting method using a subset of expressions in the whole expression set [64]. This method defines a metric to measure expressional redundancy between facial delta-blendshapes and uses the metric to find a subset of delta-blendshapes with fewer redundancies. Subsequently, the subset is used to generate a 3D facial shape by minimizing the distance between the 3D face and facial landmark points via optimization using the gradient descent algorithm. Although this method is similar to the proposed method here in that it uses the subset of less-correlated facial expressions, this does not sufficiently consider cancellations between expressional motions. Thus, this method can efficiently address the “redundancy“ between similar-shaped delta-blendshapes, making it avoid faces with exaggerated expressions. But, it cannot describe fine facial details made from composites of completely different expressions. Also, the expression selection scheme was heuristically defined in [64]. In contrast, the proposed method addresses the expressional redundancy and the facial shape details simultaneously by introducing two facial similarity metrics. The most similar expression to the target can be found with the *expressional similarity*. Subsequently, the appropriate expression coefficient is obtained with the *relative similarity*. In addition, it is demonstrated mathematically that the greedy selections in each iteration can lead to maximizing the expression similarity to the target from the global perspective.

III. EXPRESSION BLENDSHAPE FITTING

We briefly review the delta-blendshape fitting to define our notation and explain the method clearly. The facial mesh or blendshape is a set of vertices denoted by capital letters (e.g., P or F) and its component vectors, characterized by bold lowercase letters (e.g., $\mathbf{p}_i$ or $\mathbf{f}_{1,i}$). A facial blendshape mesh B can be represented as a linear combination of n_{exp} blendshape expression models $\mathbf{B} = [B_1, \ldots, B_{n_{\text{exp}}}]$ as $B = \mathbf{B}\mathbf{e}^T = \sum_{i=1}^{n_{\text{exp}}} e_i B_i$, where $\mathbf{e} = [e_1, \ldots, e_{n_{\text{exp}}}]$ is a vector of expression coefficients and $(\cdot)^T$ indicates a transposition operation [16], [39]. In the delta-blendshape approach, the facial shape is based on the following constraints to B:

- B_1 is a neutral facial shape.
- All the non-neutral facial weights are bounded between 0 and 1; that is, $0 \leq e_i \leq 1$ for $2 \leq i \leq n_{\text{exp}}$.
- The sum of all the weights is 1 so that $e_1 = 1 - \sum_{i=2}^{n_{\text{exp}}} e_i$.

By the above constraints, the blendshape mesh B can be represented in terms of the displacements from the neutral face, the delta-blendshapes $\Delta B_i = B_i - B_1$, as follows:

$$B(\mathbf{e}) = B_1 + \sum_{i=2}^{n_{\text{exp}}} e_i \Delta B_i. \tag{1}$$

Thus, $(n_{\text{exp}} - 1)$ coefficients need to be determined to describe a target facial expression. For simplicity, we use $\mathbf{e}$ to refer to all the expression coefficients except for the neutral coefficient in the following equations and discussions. A facial mesh F can be represented for rotation and translation to the facial model in (1) as follows:

$$F = \mathbf{R}\left(\mathbf{B}\mathbf{e}^T\right) + \mathbf{t}, \tag{2}$$

where $\mathbf{R} \in \mathbb{R}^{3\times 3}$ is a rotation matrix and $\mathbf{t} \in \mathbb{R}^3$ is a translation vector. Fitting an expression model is a procedure for finding an optimal $\mathbf{e}$, whereby a facial shape is generated to reveal the target's facial expression most accurately. The blendshape's expression weights are approximated by minimizing the distances between the target's facial features and the corresponding points on the 3D facial blendshape. In previous works [27], [39], 3D facial point clouds captured by 3D laser scanners or depth sensors such as Microsoft Kinect have been used to fit expression blendshapes to a target. Some other works [8], [18], [28], [60] have utilized sparse facial landmarks on an image jointly with 3D point clouds to match the target's expression more precisely or used landmarks alone to fit the blendshapes for more general applicability.

The energy term E_{pnt} to be minimized for point cloud matching can be defined as follows:

$$E_{\text{pnt}} = \frac{1}{n_{\text{ver}}} \sum_{j=1}^{n_{\text{ver}}} \left|\mathbf{f}_j - \mathbf{p}_j\right|^2, \tag{3}$$

where $\mathbf{f}_j \in \mathbb{R}^3$ is the j^{th} vertex of the facial mesh F, $\mathbf{p}_j \in \mathbb{R}^3$ is the closest point to $\mathbf{f}_j$ among the target point clouds, and n_{ver} is the number of vertices of the facial mesh.

The energy term E_{fea} for the landmark matching is defined as follows:

$$E_{\text{fea}} = \frac{1}{n_{\text{fea}}} \sum_{k=1}^{n_{\text{fea}}} \left|\Pi\left(\mathbf{f}_{v_k}\right) - \mathbf{l}_k\right|^2, \tag{4}$$

where v_k is the corresponding vertex index of the facial mesh F to the k^{th} two-dimensional (2D) facial landmark $\mathbf{l}_k \in \mathbb{R}^2$, $\Pi(\cdot) : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ is a perspective projection operator, and n_{fea} is the number of facial landmarks utilized for the fitting.

In order to avoid over-fitting noisy points and getting stuck in the local minima, an additional energy term is essential to regularize the expression coefficients or the facial shape. Several regularization terms have been used on the basis of the norms of the coefficients and the types of shape priors, which are defined on the basis of the probabilistic distributions of facial shapes in the training set [37], [39]. The $\mathbb{L}_1$-norm and the $\mathbb{L}_2$-norm of the delta-blendshape coefficients [30], [31], [32] are the regularization methods widely used for facial fitting to form stable and natural facial shapes in the delta-blendshape representation by controlling the concurrent activations of the delta expressions. The $\mathbb{L}_1$- and $\mathbb{L}_2$-coefficient regularization terms can be defined as follows:

$$\begin{aligned} E_{\mathbb{L}_1} &= \frac{1}{n_{\text{exp}}} |\mathbf{e}| = \frac{1}{n_{\text{exp}}} \sum_{i=2}^{n_{\text{exp}}} |e_i|, \\ E_{\mathbb{L}_2} &= \frac{1}{n_{\text{exp}}} |\mathbf{e}|^2 = \frac{1}{n_{\text{exp}}} \sum_{i=2}^{n_{\text{exp}}} {e_i}^2. \end{aligned} \tag{5}$$

In the delta-blendshape fitting methods, the optimal expression coefficients $\mathbf{e}^*$ can be obtained by minimizing the total fitting energy E_{fit} as follows:

$$\mathbf{e}^* = \arg\min_{\mathbf{e}} E_{\text{fit}}, \tag{6}$$

where

$$E_{\text{fit}} = \omega_{\text{pnt}} E_{\text{pnt}} + \omega_{\text{fea}} E_{\text{fea}} + \omega_{\mathbb{L}_1} E_{\mathbb{L}_1} + \omega_{\mathbb{L}_2} E_{\mathbb{L}_2}, \tag{7}$$

and $\omega_{\text{pnt}}, \omega_{\text{fea}}, \omega_{\mathbb{L}_1}$, and $\omega_{\mathbb{L}_2}$ are the constants that balance the energy terms. In practice, the geometric parameters $\mathbf{R}$ and $\mathbf{t}$ in (2) need to be determined in advance to solve the problem in (6). It is achieved by optimizing jointly or alternately on $\mathbf{R}$, $\mathbf{t}$, and $\mathbf{e}$ until the expression coefficients $\mathbf{e}$ converge to minimize the fitting energy E_{fit} in (7).

IV. SELECTIVE EXPRESSION MANIPULATION

In the delta-blendshape approach, the facial shape is given by a combination of the delta-blendshapes as defined in (1). Artists have widely used it for character modeling and user interactions since it enables the facial shape to be locally controlled and semantically interpretable. However, when this approach is used to estimate a set of expressions from a human face, the interpretability is often decreased because the weights of the expression blendshapes obtained by the fitting method are not uniquely determined. The concurrent activations of the expression models that are represented in similar facial regions can interfere with each other, decreasing the semantic meaning of each other's expression. Thus, fitting with the whole-expression blendshapes can cause the fitting procedure to become over-fitted or stuck in the local minima.

To address the problem, we attempted to find a set of expressions relevant to the target from among the entire set of expression blendshape models instead of using the whole-expression models. Inspired by greedy algorithms [22], [23], we propose a method whereby a subset of expressions is obtained by iteratively selecting a delta expression candidate that shows the highest expression similarity to the target. Figure 3 depicts an overview of the SEM, in which the six expression delta-blendshapes related to the mouth shape ΔB_i $(1 \leq i \leq 6)$ are represented for simplicity. By explaining vertex displacements as a facial motion, the expressional similarity is defined as a measure in terms of the facial movements of the blendshape model. Thus, SEM selectively obtains a set of expression blendshapes; it selects a series of facial motions required to compose the target facial shape. In order to interpret the expression delta-blendshapes as facial motions more semantically and precisely, the selected expressions should not be canceled out by the other blendshapes. As a result, SEM obtains a set of less inter-correlated expressions that are semantically close to the target face by removing expression candidates that lead to dissimilar motions from those of the target.

First, we introduce the SEM under the assumption that a target facial mesh has the same vertex correspondence to an expression blendshape. The two kinds of expressional similarities are defined by measuring the motion difference between the target mesh and the blendshape model. Based on the similarities, a robust method for predicting and selecting a subset of less inter-correlated expressions for best representing the target is described. Then, we extend the proposed method to expression selection for given landmarks and face point clouds.

A. THE SIMILARITY BETWEEN TWO MESHES

From an initial facial mesh $F_0 = [x_0, y_0, z_0, \ldots, x_{n_{\text{ver}}}, y_{n_{\text{ver}}}, z_{n_{\text{ver}}}] \in \mathbb{R}^{(3 \times n_{\text{ver}})}$, assume that two facial meshes F_1 and F_2 are generated by adding expressional motions to F_0. The motion of F_0 toward F_i $(i = 1, 2)$ is defined as $\Delta F_i = F_i - F_0$. The motion correlation C_M between F_1 and F_2 w.r.t. F_0 can be measured by as follows:

$$C_M\left(F_1, F_2\right) = \left(F_1 - F_0\right) \cdot \left(F_2 - F_0\right) = \Delta F_1 \cdot \Delta F_2. \tag{8}$$

The expressional similarity (*ESim*) is defined by normalizing the motion correlation in (8) for the magnitude of both motion vectors:

$$ESim\left(\Delta F_1, \Delta F_2\right) = \frac{\Delta F_1 \cdot \Delta F_2}{|\Delta F_1|\,|\Delta F_2|}. \tag{9}$$

Assuming that F_2 is an expressional facial shape of the target, *ESim* measures the similarity between the facial motion ΔF_1 and the facial motion of the target ΔF_2. As *ESim* measures the cosine similarity of two motions, it quantifies the normalized value of producing similar facial shapes, ranging from –1 to 1. Using *ESim*, a blendshape motion is selected in each iteration that produces a facial shape most similar to the target facial shape.

Figure 4 depicts two sets of facial shapes F_1 and F_2 generated from F_0. In Figure 4(a), they have similar poses, of which *ESim* is 0.82. Conversely, in Figure 4(b), the *ESim* is -0.81 when two faces have different poses corresponding to widely opening and closing eyes.

The motion correlation in (8) can be considered an $(n_{\text{ver}} \times 3)$-dimensional projection of ΔF_2 onto ΔF_1 or vice versa. Based on the geometric property, we introduce the relative similarity that measures the relative amount of one's expressional motion to the other. The relative similarity (*ERSim*) of ΔF_1 to ΔF_2 is defined by normalizing the motion correlation for the magnitude of ΔF_1:

$$\begin{aligned} ERSim\left(\Delta F_1, \Delta F_2\right) &= \frac{1}{|\Delta F_1|}\left(\Delta F_2 \cdot \frac{\Delta F_1}{|\Delta F_1|}\right) \\ &= \frac{\Delta F_1 \cdot \Delta F_2}{|\Delta F_1|^2}. \end{aligned} \tag{10}$$

This metric measures the magnitude of ΔF_1 relative to ΔF_2 in terms of the expressional motion. When the magnitudes of ΔF_1 and the projection of ΔF_2 onto ΔF_1 are the same, *ERSim* produces a unit value. On the other hand, when the magnitude of the two motions varies, *ERSim* provides a scale multiplier that makes ΔF_1 equal to the projection of ΔF_2, i.e., an optimal magnitude of ΔF_1 to be made similarly to ΔF_2. In short, *ESim* in (9) measures the directional similarity between the facial motions, whereas *ERSim* in (10) quantifies the similarity between the facial shapes that the facial motions change from the initial face. Thus, *ERSim* is used in SEM for two purposes: (a) measuring the difference of a facial shape from the target face from a motional perspective and (b) determining an optimal magnitude of the expression delta-blendshapes to be made similarly to the target face.

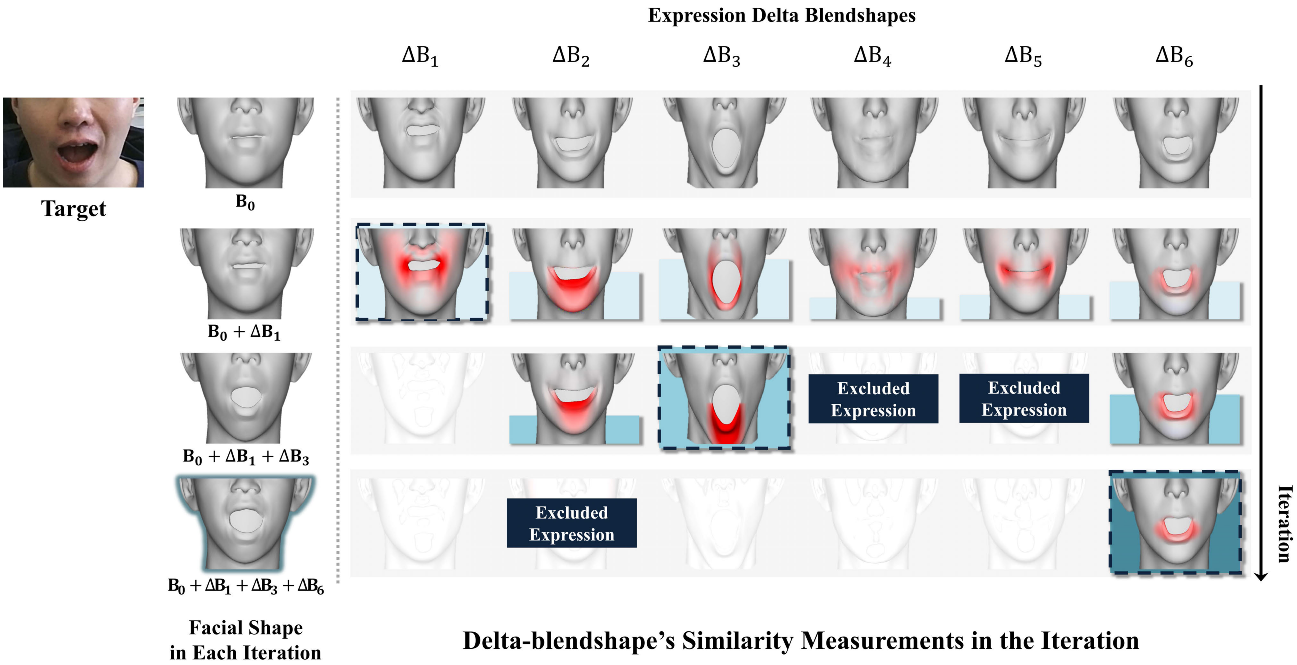
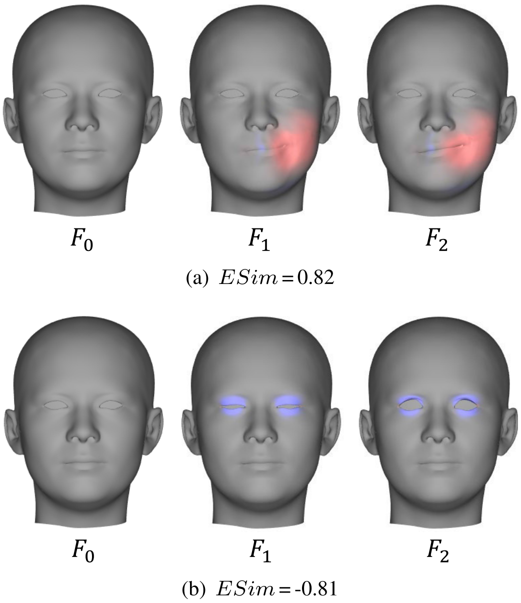


FIGURE 3. An overview of Selective Expression Manipulation, in which only the region around the mouth is depicted for clarity. Based on the expression similarity of each expression blendshape to the target (colored in red), a set of facial motions that make a target facial shape is selected sequentially while removing redundant candidates are removed to ensure unique and semantic choice.



(a) $ESim = 0.82$

(b) $ESim = \text{-}0.81$

FIGURE 4. For visualization, ***ESim*** is visualized per vertex rather than integrating values over the entire mesh and represented in red (positive) and blue (negative).

Figure 5 shows an example where *ERSim*s of ΔF_1 relative to ΔF_2 are measured by varying the magnitude of ΔF_1. As shown in Figure 5(d), *ERSim* is 1.0 when two facial motions are the most similar. Moreover, when F_1 $(= F_0 + \Delta F_1)$ has either less or more expressive shape than F_2, *ERSim* can provide a multiplier α for the facial motion ΔF_1 to make the facial shape similar to F_2, such that $F_0 + \alpha \Delta F_1 \approx F_2$ ($\alpha = 5.0$ and 0.5 for 5(c) and 5(e), respectively). Thus, for a unit motion ΔF_1, *ERSim* provides an optimal magnitude of the motion similar to ΔF_2. In SEM, the selected blendshape motion at each iteration is magnified to be close to the target facial shape from an expressional perspective by measuring *ERSim*.

B. SEM ON CORRESPONDING MESHES

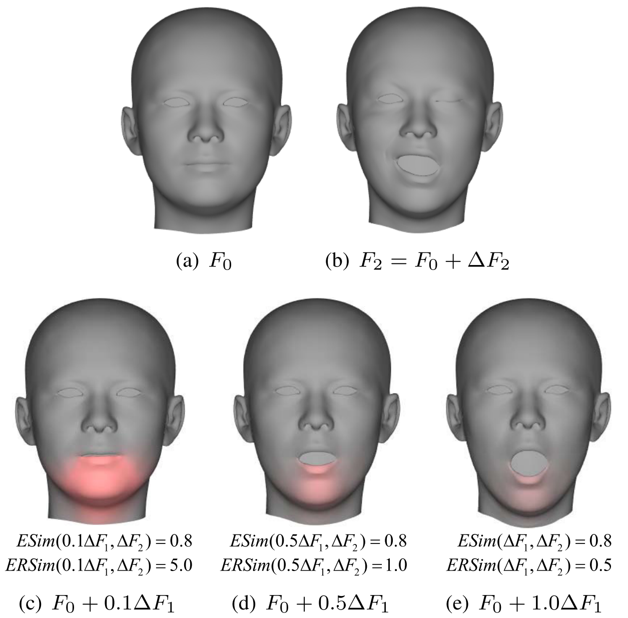
1) OBJECTIVE

Let $T \in \mathbb{R}^{(3 \times n_{\text{ver}})}$ be a target facial mesh and B be the facial shape representation using the blendshapes defined in (1). In SEM, the target face is derived from an initial face by a series of facial motions, that is, the delta-blendshapes. Then, the expression fitting energy of the blendshape model B to be minimized with respect to the expression coefficient vector $\mathbf{e}$ is defined:

$$E_{mot}(\mathbf{e}) = |(T - B(\mathbf{e})) \cdot \Delta T|, \tag{11}$$

where ΔT is the motion derived from the initial facial shape, i.e., $\Delta T = T - B_1$. The motional fitting energy in (11) can be represented in terms of the relative similarity *ERSim* in (10) as

$$\begin{aligned} E'_{mot}(\mathbf{e}) &= 1 - \sum_{i=2}^{n_{\text{exp}}} ERSim(\Delta T, e_i \Delta B_i) \\ &= 1 - ERSim\left(\Delta T, \sum_{i=2}^{n_{exp}} e_i \Delta B_i\right). \end{aligned} \tag{12}$$


$ESim(0.1\Delta F_1, \Delta F_2) = 0.8$
$ERSim(0.1\Delta F_1, \Delta F_2) = 5.0$
$ESim(0.5\Delta F_1, \Delta F_2) = 0.8$
$ERSim(0.5\Delta F_1, \Delta F_2) = 1.0$
$ESim(\Delta F_1, \Delta F_2) = 0.8$
$ERSim(\Delta F_1, \Delta F_2) = 0.5$

(a) F_0 (b) $F_2 = F_0 + \Delta F_2$

(c) $F_0 + 0.1\Delta F_1$ (d) $F_0 + 0.5\Delta F_1$ (e) $F_0 + 1.0\Delta F_1$

FIGURE 5. (a) The initial mesh F_0, (b) The target mesh F_2, (c), (d), (e) The facial shapes generated by three different magnitudes of the delta expression ΔF_1 from F_0. Each has different *ESim* and *ERSim*. *ERSim* reaches 1.0 when two expressional motions are the most similar in (d). As *ERSim* measures a relative similarity, $0.5 \cdot \Delta F_1$ is the delta expression most similar to F_2 as depicted in (d).

The full description of the derivation of (12) appears in *Appendix*. SEM determines pairs of delta-blendshapes and coefficients toward greedily minimizing (12) to satisfy the following conditions: (a) the unique set of expression blendshapes and coefficients can be obtained from a given facial mesh, (b) as few as expression coefficients can be obtained to minimize the absolute motional fitting energy $\left|E'_{mot}\right|$, and (c) the semantic meaning of each expression delta-blendshape used to represent the facial shape can be preserved.

2) PROBLEM FORMULATION

Starting from a neutral facial shape, SEM selects an expressional motion, that is, a delta-blendshape iteratively in a greedy manner. The facial shape after the t^{th} selection is defined as follows:

$$F^{t+1} = F^t + e^t \Delta B^t, \tag{13}$$

where ΔB^t and e^t are the selected delta-blendshape and its coefficient in the t^{th} iteration ($t \geq 1$), respectively. Note that the initial facial shape is a neutral facial blendshape, i.e., $F^1 = B_1$ in (1).

Since the greedy selection scheme sequentially determines the expressions one by one as it iterates, the motional energy in (12) to be minimized in each iteration t is reformulated as

$$\left(i^t, e^t\right) = \arg\min_{i, e_i} \left(1 - ERSim\left(\Delta T^t, e_i \Delta B_i\right)\right), \tag{14}$$

where $\Delta T^t = T - F^t$ and $i^t \in \{i \mid 2 \leq i \leq n_{\exp}\}$ is the expression index selected in the t^{th} iteration so that the delta-blendshape selected in the t^{th} iteration is $\Delta B^t = \Delta B_{i^t}$.

3) EXPRESSION SELECTION

As SEM iterates by (13), it determines an expression blendshape ΔB^t from among the blendshape candidates and the magnitude of the selected blendshape e^t represented for the facial shape that minimizes the fitting energy in (14). It can be accomplished by finding an expression that minimizes *ESim* in (9) and measuring its coefficient via *ERSim* in (10). The detailed description is represented in *Appendix*.

Assume that there are $(n_{\exp} - 1)$ delta-blendshape candidates $\Delta B_i \in \mathbb{R}^{(3 \times n_{\text{ver}})}$ $\left(2 \leq i \leq n_{\exp}\right)$. To determine ΔB^t, SEM selects one of the blendshape candidates most expressively similar to the target facial shape. *ESim* between the motion toward the target facial shape $\left(T - F^t\right)$ and each candidate s_i $\left(2 \leq i \leq n_{\exp}\right)$ can be measured as follows:

$$s_i = ESim\left(e_i \Delta B_i,\ \Delta T^t\right). \tag{15}$$

However, e_i, which is the magnitude of each blendshape, is unknown. As the *ESim* is defined by normalizing the magnitudes of facial motions, it is scale-invariant with a facial motion by definition in (9). Therefore, the coefficient does not affect the result of the expressional similarity as described in Figure 5. Thus, s_i in (15) can be substituted as follows:

$$s_i = ESim\left(\Delta B_i,\ \Delta T^t\right). \tag{16}$$

Once the similarities for all the candidates are measured, SEM selects the blendshape that has the largest similarity as the delta-blendshape in the t^{th} iteration ΔB^t:

$$i^t = \arg\max_i s_i, \quad \Delta B^t = \Delta B_{i^t}. \tag{17}$$

Then, the corresponding coefficient e^t is obtained using *ERSim* in (10) as follows:

$$e^t = \max(ERSim\left(\Delta B^t,\ \Delta T^t\right), 0). \tag{18}$$

The blendshape selected in the current iteration t is removed from the blendshape candidate set not to be chosen in the following iteration.

4) EXPRESSION EXCLUSION

SEM selects $(n_{\exp} - 1)$ expression blendshapes for $(n_{\exp} - 1)$ iterations in the order of expressional similarity. Expression blendshapes dissimilar to the target facial shape from a motional perspective are excluded from the set of candidates before the selection to obtain fewer expressions that are less inter-correlated. In SEM, the greedy selection at each iteration picks the most similar motion to the remaining motion required to reach the target face. Therefore, it is necessary to confirm from a global perspective that the selections of SEM iterations decrease the motional fitting energy from an initial face in (12).

We introduce an expression exclusion step to efficiently accomplish such a process using a 'one-step' backtracking strategy [24], [25], [65]. Based on the formulation in (13), we define the motional energy index from the initial face, which is measured for all the remaining candidates ΔB_i at the beginning of the t^{th} iteration:

$$r\left(t, \Delta B_i\right) = 1 - ERSim\left(\Delta T, \tilde{B}_i\right), \tag{19}$$

where $\tilde{B}_i = \sum_{k=1}^{t-1}\left(e^k \Delta B^k\right) + e_i \Delta B_i$ and e_i is calculated by (18). The term $\sum_{k=1}^{t-1}\left(e^k \Delta B^k\right)$ in $\tilde{B}_i$ is the accumulated motion using the selected expressions for $(t-1)$ iterations and the term $e_i \Delta B_i$ is the motion of the i^{th} expression among the remaining candidates. Thus, the motional energy index of the selected expression in the previous iteration, i.e., $(t-1)^{th}$ iteration, can be measured as $r\left(t-1, \Delta B^t\right)$. As a sequence of motions, an expression to be selected should not increase the absolute motional energy in (12) from the initial face. Thus, All the expression candidates that satisfy the following condition in the t^{th} iteration are excluded in the subsequent selections:

$$\left|r\left(t-1, \Delta B^t\right)\right| < \left|r\left(t, \Delta B_i\right)\right|.$$

This exclusion weeds out candidates that largely cancel out the previously selected expressions while allowing the selected motions that move toward the target. It helps SEM to non-parametrically determine the number of expressions required to represent the target expression semantically. The SEM procedure is terminated when no candidate remains by the selection and exclusion procedures.

C. SEM ON LANDMARKS

It is assumed that the target facial mesh corresponding to the blendshape model is given in Sec. IV-B. However, in real applications for blendshape fitting, the point clouds of the target face or facial landmarks on the image are given generally, as used in (3) and (4). As the vertex-point correspondence between the blendshape model and the facial landmarks is known, the expression selection on the facial landmarks is similar to the selection on the facial mesh, except for the following two conditions: 1) The facial landmarks are far fewer than the number of blendshape vertices, and 2) the facial landmarks are projected 2D points on the image, rather than 3D vertices.

When 2D facial landmarks are given, facial motion is measured on the projected plane rather than on the 3D coordinates. Let $L = [\mathbf{l}_1, \ldots, \mathbf{l}_{n_{\text{fea}}}] \in \mathbb{R}^{(2\times n_{\text{fea}})}$ and $\mathcal{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_{n_{\text{fea}}}] \in \mathbb{R}^{(3\times n_{\text{fea}})}$ be the flatten vectors of 2D facial landmarks and the corresponding vertices of the blendshape models, respectively. For the ease of comparison with the previous blendshape fitting, the same notation as in (4) is used. The (16), (18), and (19) can be modified respectively as follows:

$$s_i = ESim\left(\Pi\left(B_i\right) - \Pi\left(B_1\right), L - \Pi\left(\mathcal{F}\right)\right), \tag{20}$$

$$e^t = ERSim\left(\Pi\left(B_i\right) - \Pi\left(B_1\right), L - \Pi\left(\mathcal{F}\right)\right), \tag{21}$$

$$r\left(t, \Delta B_i\right) = 1 - ERSim\left(L - \Pi\left(F^1\right), \Pi\left(B_1 + \tilde{B}_i\right) - \Pi\left(B_1\right)\right). \tag{22}$$

SEM greedily selects expressions in the order of the expressional similarity s_i and directly computes the coefficient of the selected blendshape e^t in (21). Thus, unlike the previous blendshape-fitting methods, SEM does not require the regularization term in (7) to obtain the sparse expressions and iterative optimization methods to find the optimal coefficient for the facial landmarks. In addition, for the given geometric parameters and the target facial landmarks, SEM selects the unique set of expressions for each trial.

D. SEM ON POINT CLOUDS

For point clouds, the vertex-point correspondences to the blendshape model are not given. In the previous blendshape fitting, the closest point to each vertex of the blendshape is chosen during the optimization expressed in (6). Similarly, SEM finds the corresponding pairs using Iterative Closest Point (ICP) [66]. However, in SEM, the corresponding pairs are obtained for each blendshape. Thus, SEM finds the target point set $P_i = [\mathbf{p}_{i,1}, \ldots, \mathbf{p}_{i,n_{\text{ver}}}]$ closest to the delta-blendshape candidate ΔB_i $\left(2 \leq i \leq n_{\exp}\right)$ using ICP in each iteration of the expression selection, which can be denoted as

$$\left[P_i, e_i^*\right] = \underset{P, e_i}{\arg\min} \left\| \left(F^t + e_i \Delta B_i\right) - P \right\|^2 \tag{23}$$

where P is the point set closest to the facial mesh with a single expressional motion ΔB_i of magnitude e_i, i.e., $F^t + e_i \Delta B_i$. The objective function in (23) is to find the closest point set P_i with respect to e_i, and e_i^* is not further used. Once all the correspondences are obtained, the parameters in (16), (18), and (19) can be measured for each candidate as follows:

$$s_i = ESim\left(\Delta B_i, P_i - F^t\right),$$
$$e^t = ERSim\left(\Delta B^t, P_i - F^t\right), \tag{24}$$
$$r\left(t, \Delta B_i\right) = 1 - ERSim\left(P_i - F^1, \tilde{B}_i\right). \tag{25}$$

We assume that the expression is selected by (17) from among the candidates in (24) at the t^{th} iteration, and denote P^t as the point set of the selected expressions. As P^t is updated to include more accurate matches at each iteration, the coefficients for the previously selected expressions can be refined by using the updated correspondence.

The point cloud matching error using the closest point in (3) is largely non-linear and it has several local minima. Therefore, without strong regularizations or priors, jointly optimizing the full expressions using ICP causes the fitting procedure to get stuck in the local minima in many cases. SEM efficiently avoids these local minima and finds a better solution for the target point cloud by providing such strong priors of target expressions in the order of the motional similarity.

V. EXPERIMENTAL RESULTS

We conducted two main experiments to evaluate the performance of our proposed method. In the first part, the performance for facial expression reconstruction and alignment was evaluated against existing schemes. We categorized the fitting optimization functions used in the previous facial reconstruction methods into four baselines. The performance comparisons to the baselines were performed to verify the SEM scheme over $\mathbb{L}_1$, $\mathbb{L}_2$, and PCA regularizations. In addition, the comparisons with the state-of-the-art reconstruction methods were performed to validate the performance improvements of SEM. In the second part, we tried to evaluate the uniqueness and semantics of the proposed method by correlating the expression coefficients obtained from the expression fitting with two closely related attributes to facial expressions.

A. 3D EXPRESSION RECONSTRUCTION AND ALIGNMENT

1) QUANTITATIVE METRIC FOR FACIAL RECONSTRUCTION

To evaluate facial alignment performance, we measured the normalized error (*NE*) and the relative similarity (*ERSim*) of the obtained shape to the target shape. *NE* is the distance to the target points normalized by the inter-ocular distance. *ERSim* measures closer to one when the generated shape is more similar to the target in terms of facial motions, as defined in (13). In addition, the total redundancy among the blendshapes, denoted as *Corr*, was measured. Please refer to *Appendix*-VI for details of the measurement metrics.

2) COMPARISONS WITH THE BASELINE METHODS

The early methods for the blendshape fitting used point clouds from depth sensors or 3D scanners. By contrast, recent methods have widely used facial images or facial landmarks obtained from the images rather than using point clouds to synthesize a 3D human face. Regardless of the types of the target features used for the blendshape fitting, many previous works using the delta-blendshape models [8], [27], [28], [36], [60], [61] have been obtained the coefficients of the expression delta-blendshapes by minimizing the fitting error E_{fit} in (7) under the assumption that the user-specific blendshapes were given. Unlike 3DMM based methods, these works did not use a low-dimensional subspace decomposition for the expression blendshapes such as PCA. Instead, the original shape of each expression blendshape was intended to be preserved as possible. Thus, either $\mathbb{L}_1$ or $\mathbb{L}_2$ regularization term for the expression coefficients in (7) was adopted in these methods to prevent the facial shape from being over-fitted and from having exaggerated expressions. To cover the delta-blendshape fitting schemes used in these methods, the fitting energy of the two baseline methods with the $\mathbb{L}_1$ and $\mathbb{L}_2$ regularizations are represented in (7).

The $\mathbb{L}_1$ and $\mathbb{L}_2$ norms were mutual-exclusively used in our experiments for the reliable measurement, i.e., either $\omega_{\mathbb{L}_1}$ or $\omega_{\mathbb{L}_2}$ was zero. For simplicity, we denoted the results obtained by utilizing the $\mathbb{L}_1$ and $\mathbb{L}_2$ regularizations as $\mathbb{L}_1$Reg and $\mathbb{L}_2$Reg, respectively. The $\mathbb{L}_0$ norm, which is a counting loss for the non-zero expression coefficients, had been considered for our experiments because it has been used to make sparse representations in other applications [67]. However, the $\mathbb{L}_0$ norm produced unreliable results in our tests and it was challenging to find common constant values or an interval to balance the normalization term and the others. Therefore, we do not include the $\mathbb{L}_0$ norm in the baseline experiments. In addition to $\mathbb{L}_1$Reg and $\mathbb{L}_2$Reg, the performance comparisons were conducted with the fitting schemes using PCA-decomposed basis vectors, which have been used in many parametric model-based fitting methods [30], [49], [50], [51], [53], [54], [55], [68]. To cover the parametric model fitting schemes used in these methods, two types of expression bases were used for the baseline comparison. Firstly, we obtained the PCA basis by decomposing the entire expression set, including a neutral model. The second one was obtained by decomposing the delta shapes of the expression blendshapes [69], which has been widely used in PCA-based facial fitting approaches. We denoted the results of these methods as PCAExp and PCADel, respectively. Following the standard regularization technique for PCA coefficients [51], [53], [54], the $\mathbb{L}_2$ regularization to coefficients weighted with the inverse of PCA eigenvalues was used, which is the so-called “Mahalnobis distance” regularization [70].

a: RESULTS ON FACIAL LANDMARKS

The fitting energy for the facial landmarks in (7) can be represented with $\omega_{pnt} = 0$ to disable the point cloud loss. To obtain an accurate neutral face model for each facial expression image for composing the user-specific blendshapes [36], we used the Extended Cohn-Kanade (CK+) database [21], where each sequence begins at the neutral expression and ends at the peak expression. For 593 sequences from 12 subjects who portrayed seven basic emotions in the CK+ database, the images in the first frame were used to obtain the user identities. Then, SEM, $\mathbb{L}_1$Reg, and $\mathbb{L}_2$Reg used the obtained user-specific blendshape set directly. In contrast, we decomposed the obtained blendshape set into the PCA spaces for PCAExp and PCADel. Among the 68 landmark points annotated in the CK+ database images, it is difficult to clearly distinguish between the inner and outer lip corners in practice. In addition, the facial boundary landmark points can be defined ambiguously depending on the facial pose. Therefore, we used 49 landmarks for reliable evaluation by excluding 2 inner lip corners and 17 facial boundary points. For the performance evaluation, we split the 49 inner landmarks of the CK+ database into two sets as represented in the first column of Figure 6: *used* and *unused* points. One is for the facial fitting and the other is for measuring facial alignment accuracy at unobserved points. For a fair comparison, the non-linear optimization for all the methods in the baseline experiments was performed using an off-the-shelf BFGS optimizer [71], [72] to constrain the boundaries of the expression coefficients.

The distance errors of the expression fitting methods on the CK+ database are summarized in Table 1. *NE* is measured for *all* (49*pts*), *unused* (26*pts*), and *used* (23*pts*) landmarks. The errors between the baseline methods do not show

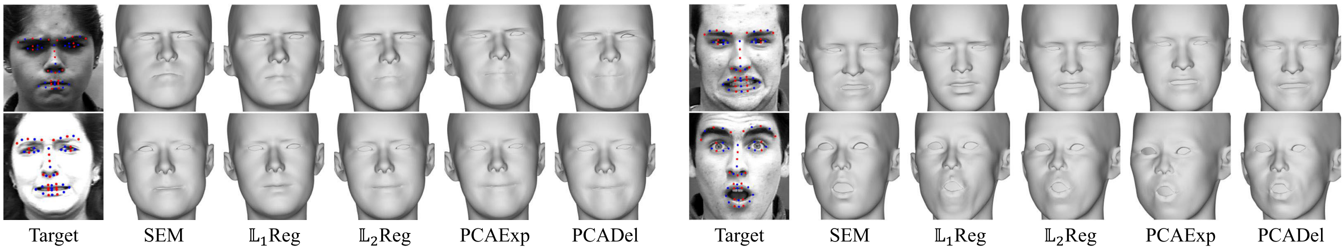


FIGURE 6. 3D Facial shapes obtained by fitting 23 of the 49 inner landmarks on the faces of the CK+ database. The 23 *used* points are marked in red. Best viewed in zoom-in.

TABLE 1. *NE* for *All* (49*pts*), *Unused* (26*pts*), and *Used* (23*pts*) Landmarks on the CK+ database.

Measurement		SEM	**Baseline**			
Metric	**Target**		$\mathbb{L}_1$Reg	$\mathbb{L}_2$Reg	PCAExp	PCADel
NE	*All*	0.043	0.056	0.056	0.057	0.056
	Unused	0.044	0.060	0.060	0.058	0.061
	Used	0.043	0.050	0.054	0.051	0.050

TABLE 2. *ERSim* and *Corr* measurements for the facial shape fitted on facial landmarks.

Metric	SEM	**Baseline**			
		$\mathbb{L}_1$Reg	$\mathbb{L}_2$Reg	PCAExp	PCADel
$ERSim$	0.764	0.614	0.632	0.619	0.623
$Corr$	0.038	0.088	0.086	0.110	0.101

TABLE 3. *Corr* measurements for positive and negative motion correlations.

Metric	SEM	**Baseline**			
		$\mathbb{L}_1$Reg	$\mathbb{L}_2$Reg	PCAExp	PCADel
$Corr(+)$	0.032	0.038	0.039	0.048	0.036
$Corr(-)$	0.006	0.050	0.047	0.062	0.065

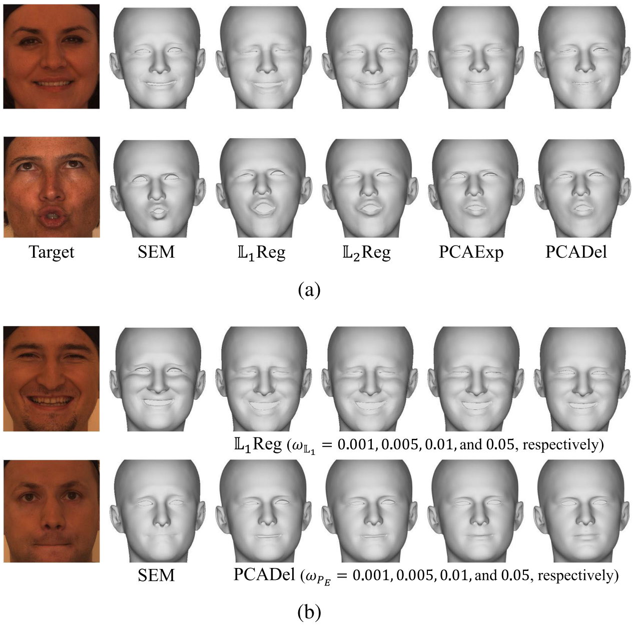

Target
SEM
$\mathbb{L}_1$Reg
$\mathbb{L}_2$Reg
PCAExp
PCADel
(a)
$\mathbb{L}_1$Reg ($\omega_{\mathbb{L}_1}$ = 0.001, 0.005, 0.01, and 0.05, respectively)
SEM
PCADel (ω_{P_E} = 0.001, 0.005, 0.01, and 0.05, respectively)
(b)

FIGURE 7. Facial shapes obtained from SEM and the baseline methods (a) without and (b) with regularization on the Bosphorus database [73]. Best viewed in zoom-in.

meaningful differences. However, the results establish that SEM outperforms the baseline methods in facial alignment. In particular, SEM shows a significantly higher performance gain when it is evaluated for the *unused* landmarks. In contrast, the baseline methods do not accurately predict the *unused* landmarks compared to *used* ones, leading to an increase in the *NE* measured for *all* landmarks. Figure 6 visualizes the facial shapes obtained from these methods. SEM produces expressive faces with the expressions visually unique to the baseline methods. In contrast, the baseline methods generate exaggeratedly expressive faces in some facial images or less-expressive faces in others, even with the same regularization balancing constants.

The measurements of *ERSim* and *Corr* are summarized in Table 2. The significantly lower *Corr* and higher *ERSim* than the baseline methods demonstrate that SEM can represent the target more expressively and semantically with significantly less correlation between the delta-blendshapes. In contrast, the baseline methods combine all the delta expressions to make the facial points of the blendshapes as close as possible to the given facial landmarks without knowing the appropriate expressions to be matched to the target expression. As a result, it may cause the obtained expressions to become redundant and higher *Corr* measurements (i.e., more than two times higher than SEM), implying that the delta shapes composed of the facial shape are motionally redundant.

TABLE 4. Quantitative performance comparisons with baseline methods on the Bosphorus database.

Metric	SEM	Baseline			
		$\mathbb{L}_1$Reg	$\mathbb{L}_2$Reg	PCAExp	PCADel
NE	0.0427	0.0567	0.0536	0.0580	0.0579
$ERSim$	0.846	0.639	0.655	0.599	0.602
$Corr$	0.830	4.504	4.827	5.450	5.368

For examining the redundancy between the obtained blendshapes in more detail, *Corr* is measured for positive and negative parts, which are $e_{F,i}e_{F,j}\left(\Delta B_i \cdot \Delta B_j\right) > 0$ and < 0 of (33) in *Appendix*. A higher $Corr(+)$ means that the expression blendshapes have redundant motions and a higher $Corr(-)$ means that the expressions tend to cancel out the others. In other words, the lower $Corr(+)$ and $Corr(-)$ can represent better target expressions. Table 3 summarizes the correlation measurements. SEM obtains expressions with lower $Corr(+)$ and $Corr(-)$ than the baseline methods. Cancellations between the expressions can cause each blendshape not to represent its own facial shape, considerably decreasing the semantics of the blendshapes. Compared to the baseline methods, the notably lower $Corr(-)$ value demonstrates that SEM can produce a set of expressions semantically by avoiding the cancellations between the blendshapes.

b: RESULTS ON 3D POINT CLOUDS

For the 3D point clouds, the Bosphorus database [73] was used. This database captured the 3D scans of 4,666 scans from 105 subjects. We used 2,603 of the 4,666 scans for the performance evaluation after excluding facial scans without expressions. The user-specific expression blendshapes were obtained from the scans of the neutral face in the Bosphorus database. The Bosphorus database provided approximately 30,000 valid facial points per scan. The point clouds were randomly sampled 10 times to produce sets for the performance evaluation. For evaluations on point clouds, the facial landmark term was not used. Thus, the fitting energy of the baseline methods can be represented as E_{fit} in (7) with $\omega_{fea} = 0$.

The average performances for the methods are summarized in Table 4. Figure 7(a) shows the visualizations of the facial shapes obtained from the proposed and the baseline methods. As strong constraints such as the facial landmarks are not used in this experiment, the facial shapes of the baseline methods tend to be trapped in the local minima as shown in Figure 7(a). PCA helps prevent the facial shapes from being overshot into local minima by decomposing the basis to move globally. However, it is shown that the globally moving basis often results in missing local details. In contrast, SEM reliably finds the target expression by capturing the distinct expressions in the order of the motional similarity, showing the lowest *NE* and the highest *ERSim*, and considerably lower *Corr* measurements than the baseline methods.

Figure 7(b) shows the facial shapes obtained from the baseline methods by using different regularization weights. It is shown that regularization does not help to form expressions with different shapes. Instead, it decreases the overall expressions of the facial shape and increases the target error as the magnitude of the weight increases. The baseline approaches do not often produce much-exaggerated facial shapes even when no regularization is used; because a sufficient number of fitting points are provided. Instead, the method often fails to find or mis-capture the target expressions. The cancellation between expressions (higher *Corr*) allows the baseline methods to generate various shapes, increasing the possibility of overshooting or getting stuck into the local minima. Also, the cancellation significantly decreases the semantics of the blendshapes, being unable to sufficiently utilize the strong priors of the blendshape faces or motions. Finally, it is demonstrated that SEM reliably constructs a realistic face highly correlated with the target facial motion (higher *ERSim*) with a combination of less inter-correlated expressions (lower *Corr*).

3) COMPARISONS WITH THE STATE-OF-THE-ART METHODS

In this section, we compared the results quantitatively and qualitatively with the state-of-the-art facial reconstruction methods using the 3D facial model to validate the performance of the proposed method. For thorough validations, multiple facial reconstruction methods recently proposed were used for the comparisons: ones proposed in the works of *Cao16* [9], *Zhu16* [50], *Chang18* [54], *Sanyal19* [68], *Guo20* [55], and *Kang20* [64], respectively. All of these methods align the 3D facial models [17], [49], [75] to a facial image and produce matching landmark points. SEM, *Cao16*, and *Kang20* required the facial landmarks for aligning the model to the face to obtain the facial shape from an image, whereas the methods of *Chang18*, *Sanyal19*, and *Guo20* needed the target image and a face bounding box on the image. For providing approximations of the facial landmarks and the bounding box in a target image, we used the Supervised Descent Method (SDM) [76], [77], which was trained with the 300W database [78]. The results of SDM, which has been still used widely today for facial alignment tasks, were also included in the quantitative comparison. We used two public databases of facial emotional expressions to validate the facial fitting performance on various types of expressions: CK+ [21] and AffectNet [74] databases. In the CK+ database, facial images were captured indoors under constrained conditions, where subjects were requested to pose specific emotional expressions. Therefore, it enables balanced evaluations of basic expressions. In contrast, in the AffectNet database, in-the-wild facial images were collected online, allowing the methods to be evaluated extensively on natural scenes. We used 10 uniformly sampled images per sequence in the CK+ database for evaluation, amounting to 5,930 facial images. The AffectNet database comprises 1,000K facial images with 68 facial landmark

TABLE 5. Performance comparisons with state-of-the-art methods on the CK+ and AffectNet databases.

Database	*SEM*	*SDM*	*Cao16*	*Zhu16*	*Chang18*	*Sanyal19*	*Guo20*	*Kang20*
CK+	0.0595	0.0745	0.0948	0.1083	0.1010	0.1211	0.0925	0.0652
AffectNet	0.0936	0.1443	0.1239	0.1410	0.1324	0.1286	0.1283	0.1021
Mean	0.0892	0.1353	0.1201	0.1368	0.0729	0.0771	0.0945	0.0981

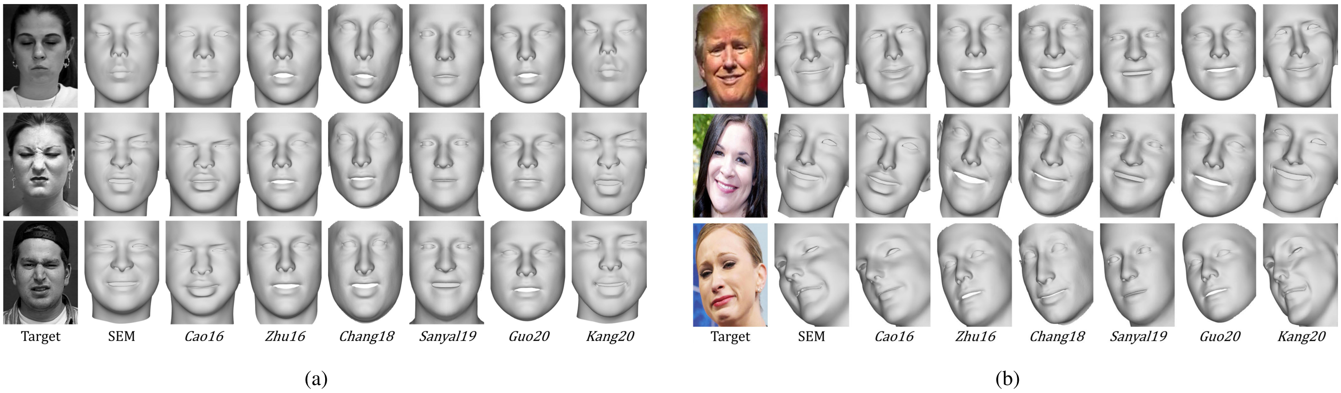


FIGURE 8. 3D Facial shapes obtained by recent fitting methods on (a) CK+ [21] and (b) AffectNet [74] database images. Best viewed in zoom-in.

annotations, which is the same as the annotations used in the CK+ database. In the AffectNet database, the landmarks of 420,300 images were annotated manually and those of the other images were obtained automatically using a facial alignment algorithm [79]. For reliable evaluation, 40,000 images in the manually annotated image set were used. We used 49 inner landmarks of 68 landmark annotations for evaluation.

The quantitative measurements of *NE* on the CK+ and AffectNet databases are summarized in Table 5. All methods show better performance on the CK+ database than the AffectNet database since in-the-wild expressions of the AffectNet database have significantly more variations. Furthermore, the images of the CK+ database have clean backgrounds and almost-frontal faces, allowing SDM to obtain better performance than facial model-based methods except for SEM. In contrast, facial model-based methods, including SEM, obtain the target landmarks with lower errors on the AffectNet database than SDM thanks to the human facial shape priors of the facial model. SEM obtains the target expression by fitting on the landmarks estimated by the landmark detector. Nevertheless, the results show that SEM decreases the mean error remarkably compared to that of SDM, enabling SEM to outperform the state-of-the-art methods in terms of the normalized distance error. This demonstrates that the increased semantic meaning of the expression blendshapes helps better characterize the facial shape and expression by providing strong priors of human expressional shapes. Figures 8(a) and 8(b) visualize facial models reconstructed by the methods on the CK+ and AffectNet databases. It can be seen that the methods of *Cao16*, *Zhu16*, *Chang18*, *Sanyal19*, *Guo20*, and *Kang20* find coarse-scale details of the target expressions sufficiently. However, they are prone to miss out on expressive details, especially around the eyes, cheeks, and mouth. In contrast, the results show that SEM reconstructs expressive faces with fine details with less deviation, demonstrating facial expression blendshapes can be semantically manipulated with SEM to find appropriate and unique target expressions.

B. SEMANTICS AND UNIQUENESS OF EXPRESSION COEFFICIENTS

In this section, we tried to evaluate the semantics and uniqueness of the proposed method. However, it is not easy to measure the performance of the expression fitting in terms of semantics and uniqueness. Therefore, we verified the proposed approach by examining the correlation of the coefficients obtained from the expression fitting with two closely related attributes to facial expressions: the Facial Action Unit (FAU) and the facial emotion. It is expected that the expression coefficients and the attributes can be highly correlated if the unique coefficients with semantics are obtained according to the FAU and emotion. Therefore, we used the Support Vector Machine (SVM) with Radial Basis Function (RBF) kernel to non-linearly correlating the coefficients with the attributes. Although it is insufficient to achieve extremely high correlations using the coefficients only, relative comparisons between the facial fitting methods allow us to verify improvements in semantics and uniqueness. The comparisons were conducted with four facial fitting methods [9], [54], [55], [68] employed in Sec. V-A3.

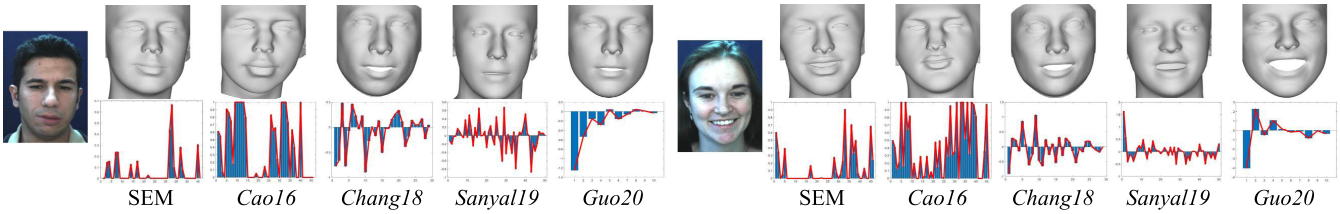


FIGURE 9. 3D facial shapes and expression coefficients for two faces with different combinations of FAUs in the DISFA database. A neutral face in the first row has two FAUs (25 and 26) and a smiling face in the second row has three FAUs (6, 25, and 26). SEM produces the expression coefficients uniquely over FAUs 25+26, whereas the other facial fitting methods produce quite different expression coefficients between the two faces.

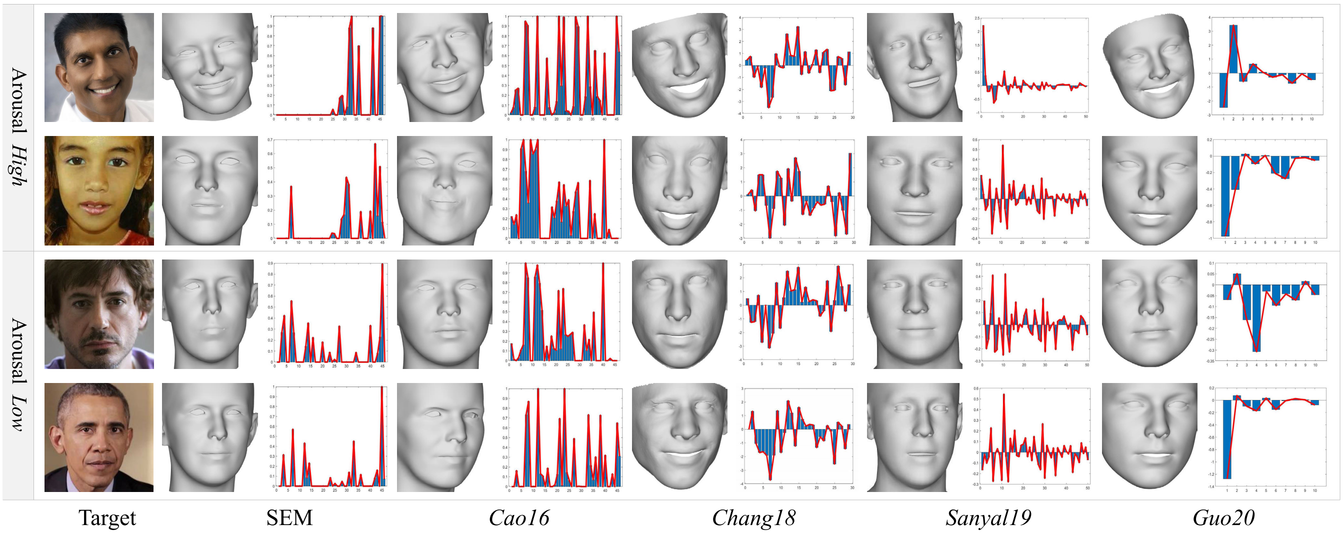


FIGURE 10. 3D facial shapes and expression coefficients for *low* and *high* Arousal faces in the AffectNet database.

1) FACIAL ACTION UNIT DETECTION

We used the Denver Intensity of Spontaneous Facial Actions (DISFA) database for the FAU detection experiment [80]. In the DISFA database, 4-minute videos of 27 subjects watching an emotional video stimulus were recorded, producing 4,845 image frames per subject. The intensities of 12 FAUs were manually coded in each frame. We used 80% of the images for training and 20% for testing the SVM classifier. Moreover, the coefficients obtained from the comparison methods were fed as inputs to the SVM and the FAUs activations were used as labels. The three standard metrics were used for the evaluation with 10-fold validation: the Mean Squared Error (MSE), the Squared Correlation Coefficient (SCC), and the classification accuracy.

Table 6 summarizes quantitative performance measurements of the comparison methods for FAU detection. The results show that SEM outperforms the previous facial fitting methods. In particular, the SCC and accuracy of SEM are measured significantly higher than the others. One of the major factors leading to the remarkably higher correlation with FAU is the unique and semantic coefficients of expressions. Figure 9 depicts 3D facial shapes and expression coefficients for two faces in the DISFA database. The first face is nearly neutral and the other is smiling. Although these faces have different expressions, both faces share two facial actions, which are "lips part" and "jaw drop" (FAU 25+26). SEM shows similar coefficient distributions between the two faces. Specifically, the coefficient distribution of the second face almost covers that of the first one, implying SEM can uniquely produce the expression coefficients for the FAUs 25+26. In contrast, the other facial fitting methods produce considerably different coefficients between the faces. The results demonstrate that SEM significantly increases the semantics and uniqueness of the expression fitting.

2) FACIAL EMOTION ESTIMATION

We used the AffectNet database [74] to estimate facial emotion values. In the AffectNet database, the facial intensity is defined in the Arousal-Valence space [81]. Since the intensity $\boldsymbol{i}$ for Arousal and Valence is composed of real numbers between -2 and 1 in this database, we divided the intensity values into 3 parts for classification: *high* for $0 \le i \le 1$, *middle* for $-1 \le i < 0$, and *low* for $-2 \le i < -1$. Thus, three FAU labels (0, 1, and 2) were used on each Arousal and Valence part to train the SVM classifiers.

Table 7 summarizes quantitative measurements of emotion estimation performance in terms of MSE and accuracy. The expression coefficients of SEM show significantly higher

TABLE 6. Quantitative performance comparisons of facial action unit detection on the DISFA database.

	SEM	*Cao16*	*Chang18*	*Sanyal19*	*Guo20*
MSE	0.6181	0.8277	1.1078	0.8621	1.0197
SCC	0.3635	0.1969	0.1046	0.1659	0.1386
Accuracy	69.12%	45.17%	32.26%	41.80%	45.17%

TABLE 7. Quantitative performance comparisons of facial emotion estimation on the AffectNet database.

		SEM	*Cao16*	*Chang18*	*Sanyal19*	*Guo20*
Arousal	**MSE**	0.4104	0.6281	0.7291	0.6803	0.6023
	Accuracy	70.17%	51.47%	50.91%	43.19%	51.98%
Valence	**MSE**	0.4227	0.6891	0.8165	0.7415	0.6515
	Accuracy	68.94%	45.43%	46.53%	40.00%	51.08%

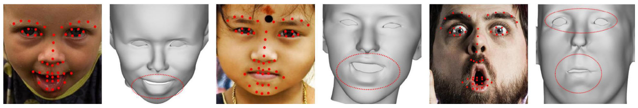


FIGURE 11. Failure cases of SEM because of the misaligned points from the landmark detector. The landmarks obtained by the detector are marked as red dots on the faces in the images.

correlations with facial emotion than those of the other reconstruction methods, which are similar results to those of FAU detection in Sec. V-B1.

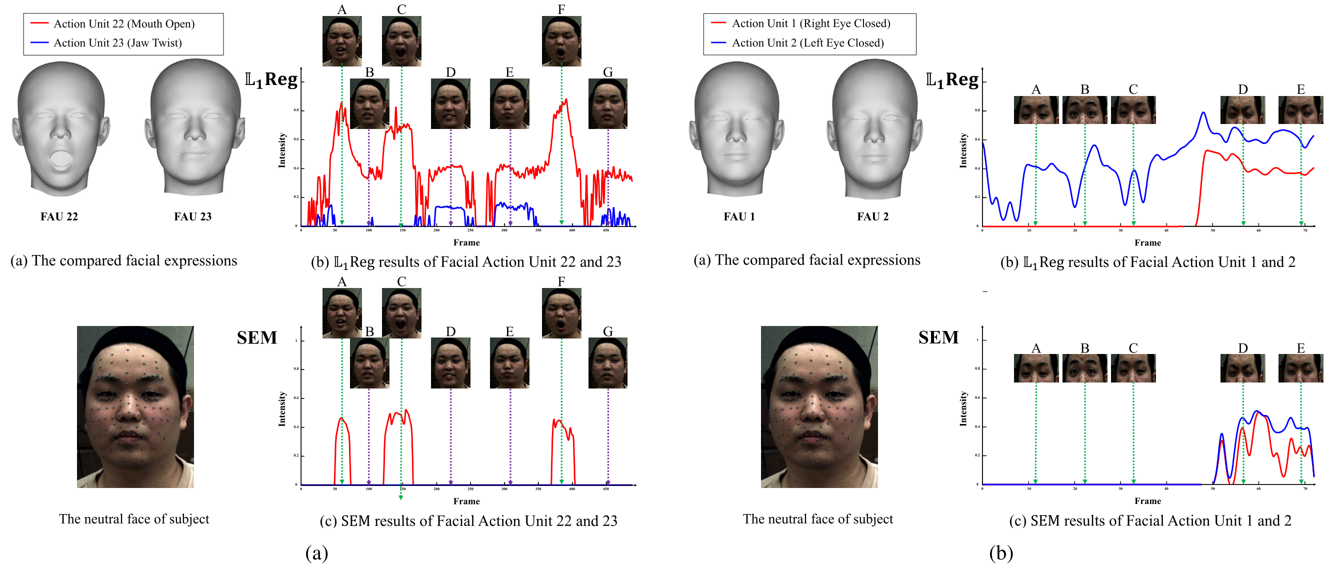
Figure 10 depicts the 3D facial shapes belonging to *high* and *low* labels in the Arousal coordinate and the corresponding expression coefficients. Each set of *high* and *low* Arousal faces represents similar emotional expressions. It is shown that SEM produces clearly distinguishable distributions of expression coefficients between intra-label and inter-label faces. In contrast, it is not easy to find common characteristics from the coefficient distributions of the other methods to classify emotions. In the experiments, the unique coefficients according to emotion enable SEM to accomplish significant improvements over the state-of-the-art fitting methods, demonstrating the benefit of SEM in terms of uniqueness and semantics.

C. STABILITY OF BLENDSHAPE SELECTION

Here, we show temporal fitting results to verify the blendshape selection stability of the proposed method. After fitting the blendshape model on facial expression sequences separately in time, we evaluated the temporal fitting stability with intensity changes for "key" expression delta-blendshapes in time. A sufficient number of accurately registered 3D facial landmark points are essential to validate the stability of blendshape selection. Thus, following the work in [38], we captured temporal facial sequences with several expressions using 108 infrared (IR) markers with multiple IR cameras to obtain highly accurate and reliable facial landmark positions in time. For the evaluation, we compared the temporal fitting stability of the proposed method with those of the $\mathbb{L}_1$ regularization method, which showed the least alignment error among the baseline methods in Section V-A2. In short, the delta-blendshape models were fitted on the temporal sequences using SEM and $\mathbb{L}_1$Reg methods, respectively. For the stability experiments, we used the same balancing weight values to those used in the baseline experiments for the fitting methods. Then, we analyzed the fitted delta-blendshape's intensities in time. We chose the two pairs of expressions in the delta-blendshapes associated with the mouth and eyes for the evaluation.

Figure 12(a) shows the intensities' changes of delta-blendshapes in time for the two key expressions, FAU 22 (mouth open) and FAU 23 (jaw twist). Although both delta-blendshapes change mouth shapes, the two expressions look clearly different from a human point of view. Thus, these expressions have quite different semantic meanings. At points A, C, and F in Figure 12(a), the subject opened his mouth largely. The results show that the previous scheme detects FAU 22 (mouth open) unevenly. Also, it recognizes FAU 23 (jaw twist) in frames that have different expressions semantically. In contrast, the proposed scheme finds mouth openings in appropriate frames without activating irrelevant expression blendshapes.

Figure 12(b) shows another example for the two key expressions, FAU 1 (right eye closed) and FAU 2 (left eye closed). The subject moves eyebrows slightly (points A-C) and closes his eyes at the end of the sequence (points D-E). Especially in the first three intervals (A-C), the subject moves their eyebrows but the eyes do not get smaller than the first frame of the sequence. Nevertheless, $\mathbb{L}_1$Reg used the expression delta-blendshapes with closed eyes (FAU 1 or FAU 2) to construct the target face for the interval. By contrast, SEM robustly distinguishes between expressions with moving eyebrows and eyes closed. These results demonstrate that the proposed scheme accurately selects the



(a) The compared facial expressions

(b) $\mathbb{L}_1$Reg results of Facial Action Unit 22 and 23

The neutral face of subject

(c) SEM results of Facial Action Unit 22 and 23

(a) The compared facial expressions

(b) $\mathbb{L}_1$Reg results of Facial Action Unit 1 and 2

The neutral face of subject

(c) SEM results of Facial Action Unit 1 and 2

FIGURE 12. The temporal fitting stability comparison between SEM and $\mathbb{L}_1$ Reg, where the intensities of two key expression delta-blendshapes associated with (a) the mouth and (b) eyes are depicted in time.

target expression blendshapes with stability and semantics in time.

D. LIMITATION AND FURTHER WORK

The previous results show that SEM can detect the detailed target expressional shapes and poses using the inner landmarks. As a result, SEM produces the 3D facial shapes of the targets more expressively in a non-parametric manner for the given facial points. Nevertheless, as SEM is inherently affected by the landmark detector, the facial expression different from the target can be obtained because of the wrong landmark positions, as described in Figure 11. This limitation could be resolved by extending the fitting procedure of SEM to the selection of optimal landmarks. In addition, it is shown that utilizing a few less inter-correlated expressions can help represent a facial expression more accurately and semantically, leading to discontinuous expression sets for a continuous image sequence. The discontinuity can be applied to find the peak expressions or the anchor frames in the sequence. Therefore, SEM can be extended to the temporal domain by propagating the other frames from the reference frames [82], [83]. The expression similarity also can be used to measure the correspondences between partially overlapped point clouds [84]. We intend to continue our research in these directions in the future.

VI. CONCLUSION

We introduced the selective manipulation approach for expression delta-blendshapes, where a target expression was approximated by a series of facial motions under the assumption that each delta-blendshape is a facial movement with semantic meaning. A series of facial motions required to compose the target facial shape were selected sequentially based on the expression similarities that quantify the motional correlation between the delta-blendshape and the remaining motion to reach the target points. The proposed method significantly decreased the cancellations between the selected expressions by excluding expression candidates that caused the motion to be uncorrelated to the target motion from an integrated perspective. Consequently, less inter-correlated expressions with considerably larger target correlations could be obtained by using the SEM. The experimental results on the public databases showed an increase in the quality, uniqueness, and semantics of the facial shape representation over the baseline and the state-of-the-art methods.

APPENDIX MOTIONAL SHAPE FITTING ENERGY AND *ERSim*

The motional shape fitting energy to be minimized in SEM can be defined as follows:

$$\begin{aligned} E_{\text{mot}} &= (P - B) \cdot \Delta T \\ &= \left(P - \left(B_1 + \sum_{i=2}^{n_{\text{exp}}} e_i \Delta B_i\right)\right) \cdot \Delta T \\ &= \left(\Delta T - \sum_{i=2}^{n_{\text{exp}}} e_i \Delta B_i\right) \cdot \Delta T \\ &= |\Delta T|^2 - \sum_{i=2}^{n_{\text{exp}}} e_i \Delta B_i \cdot \Delta T. \end{aligned}$$

Since $|\Delta T|^2$ is a constant value for the given target mesh P, we can minimize the energy $E'_{mot} = \left(E_{mot} / |\Delta T|^2\right)$ instead of E_{mot} such that

$$E'_{\text{mot}} = \frac{E_{mot}}{|\Delta T|^2} = 1 - \sum_{i=2}^{n_{\text{exp}}} \frac{(e_i \Delta B_i) \cdot \Delta T}{|\Delta T|^2}$$

$$= 1 - \frac{\left(\sum_{i=2}^{n_{\text{exp}}} e_i \Delta B_i\right) \cdot \Delta T}{|\Delta T|^2}$$
$$= 1 - \sum_{i=2}^{n_{\text{exp}}} ERSim\left(\Delta T, e_i \Delta B_i\right)$$
$$= 1 - ERSim\left(\Delta T, \sum_{i=2}^{n_{\text{exp}}} e_i \Delta B_i\right).$$

APPENDIX EXPRESSION SELECTION USING *ESim* AND *ERSim*

ESim and *ERSim* in Equations (9) and (10) defined in the original manuscript can be represented as follows:

$$\begin{aligned} ESim\left(\Delta F_1, \Delta F_2\right) &= \frac{\Delta F_1 \cdot \Delta F_2}{|\Delta F_1|\,|\Delta F_2|} \\ &= \frac{|\Delta F_1|\,|\Delta F_2| \cos\alpha}{|\Delta F_1|\,|\Delta F_2|} \\ &= \cos\alpha, \\ ERSim\left(\Delta F_1, \Delta F_2\right) &= \frac{\Delta F_1 \cdot \Delta F_2}{|\Delta F_1|^2} = \frac{|\Delta F_2| \cos\alpha}{|\Delta F_1|}, \end{aligned} \tag{26}$$

where α is the angle between the motions ΔF_1 and ΔF_2 in $(n_{\text{ver}} \times 3)$-dimensional coordinates.

Then, the measurement for s_i in Equation (17) defined in the original manuscript can be expressed as follows:

$$ESim\left(\Delta B_i, \Delta T^t\right) = \cos\alpha_i,$$

where α_i is the $(n_{\text{ver}} \times 3)$-dimensional angle between the i^{th} delta-blendshape ΔB_i and ΔT^t. Thus, finding a candidate that has the highest *ESim* in Equation (18) defined in the original manuscript is a procedure that selects an expression with the highest directional correspondence to the target motion as:

$$\alpha^t = \arg\max_{\alpha_i} \left(\cos\alpha_i\right), \tag{27}$$

where α^t is the multi-dimensional angle of the selected expression in the t^{th} iteration. The motional energy in Equation (15) defined in the original manuscript can be formulated using the selected delta-blendshape and its coefficients in Equations (17) and (18) in the manuscript, ΔB^t and e^t as:

$$\begin{aligned} E &= 1 - ERSim\left(\Delta T^t, e^t \Delta B^t\right) \\ &= 1 - \frac{\left|\Delta B^t\right| \cos\alpha^t}{\left|\Delta T^t\right|} \cdot e^t \\ &= 1 - \frac{\left|\Delta B^t\right| \cos\alpha_t}{\left|\Delta T^t\right|} \cdot \frac{\left|\Delta T^t\right| \cos\alpha^t}{\left|\Delta B^t\right|} \\ &= 1 - \cos^2\alpha^t, \end{aligned} \tag{28}$$

where $e^t = ERSim\left(\Delta B^t, \Delta T^t\right) = \frac{|\Delta T^t| \cos\alpha^t}{|\Delta B^t|}$ in Equation (18) defined in the original manuscript. As α^t in Equation (27) minimizes E in Equation (28) for $\cos\alpha^t \geq 0$, it is verified that the expression selection using *ESim* in each iteration can find an expressional motion closest to ΔT^t among the candidates and $(1 - \cos^2\alpha^t)$ is the remaining error between the motions.

APPENDIX QUANTITATIVE METRIC FOR FACIAL ALIGNMENT

The alignment error for the predicted facial shapes can be defined by the distance to the target points. Let $L_T = [\mathbf{l}_1, \ldots, \mathbf{l}_{n_{\text{fea}}}]$ be the target landmarks and S_T denote the target point clouds. Similar to the formulations in Equations (3) and (4) defined in the original manuscript, the average distances to the facial mesh F from L_T and S_T, respectively, can be calculated as follows:

$$d_{\text{lnd}}\left(F, L_T\right) = \left(\frac{1}{n_{\text{lnd}}} \sum_{j=1}^{n_{\text{lnd}}} \left|\Pi\left(\mathbf{f}_{v_j}\right) - \mathbf{l}_j\right|^2\right)^{1/2}, \tag{29}$$

$$d_{\text{pnt}}\left(F, S_T\right) = \left(\frac{1}{n_{\text{pnt}}} \sum_{k=1}^{n_{\text{pnt}}} \left|\mathbf{f}_k - \mathbf{p}_k\right|^2\right)^{1/2}, \tag{30}$$

where v_j is the vertex index of F corresponding to j^{th} landmark $\mathbf{l}_j$, $\mathbf{p}_k$ is the closest point to $\mathbf{f}_k \in F$ among S_T, and $n_{\text{lnd}}/n_{\text{pnt}}$ are the numbers of the target landmarks and the point clouds for the evaluation, respectively. As the magnitudes of d_{lnd} and d_{pnt} are scale-variant to the landmark L_T and the point set S_T of the target, an error metric used for the evaluation is defined by normalizing the distance error by the inter-ocular distance. The normalized error (NE) is measured as follows:

$$NE\left(F, L_T\right) = \frac{1}{\left|\mathbf{l}_{I_1} - \mathbf{l}_{I_2}\right|} d_{\text{lnd}}\left(F, L_T\right), \tag{31}$$

$$NE\left(F, S_T\right) = \frac{1}{\left|\mathbf{f}_{I_3} - \mathbf{f}_{I_4}\right|} d_{\text{pnt}}\left(F, S_T\right), \tag{32}$$

where I_i $(1 \leq i \leq 4)$ are the indices of the inner points of eyes for calculating the inter-ocular distance. Two more metrics were utilized for measuring the motional similarity and the redundancies between the obtained expression blendshapes. *ERSim* in the latter part of Equation (13) defined in the original manuscript was utilized to evaluate the facial shape obtained for the landmarks and the point clouds as $ERSim\left(L_T - \Pi\left(B_1\right), \Pi\left(F\right) - \Pi\left(B_1\right)\right)$ and $ERSim\left(S_T - B_1, F - B_1\right)$, respectively.

Let $\mathbf{e}_F = [e_{F,1}, \ldots, e_{F,n_{\text{exp}}}]$ be the blendshape coefficients of F. The correlation between the blendshapes $(Corr)$ can be measured as follows:

$$Corr(\mathbf{e}_F) = \sum_{i=2}^{(n_{\text{exp}}-1)} \sum_{j=(i+1)}^{n_{\text{exp}}} \left|e_{F,i} e_{F,j} \left(\Delta B_i \cdot \Delta B_j\right)\right|. \tag{33}$$

The sum of the absolute correlation in Equation (33) is measured for quantifying the total redundancy among the blendshapes used to represent the facial shape.

REFERENCES

[1] J. I. Kim, S. Li, X. Chen, C. Keung, M. Suh, and T. W. Kim, "Evaluation framework for BIM-based VR applications in design phase," *J. Comput. Des. Eng.*, vol. 8, no. 3, pp. 910–922, May 2021.

[2] I. K. Kazmi, L. You, and J. J. Zhang, ''A hybrid approach for character modeling using geometric primitives and shape-from-shading algorithm,'' *J. Comput. Des. Eng.*, vol. 3, no. 2, pp. 121–131, Apr. 2016.

[3] C. Yan, L. Meng, L. Li, J. Zhang, Z. Wang, J. Yin, J. Zhang, Y. Sun, and B. Zheng, ''Age-invariant face recognition by multi-feature fusionand decomposition with self-attention,'' *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 18, no. 1s, pp. 1–18, Feb. 2022, doi: 10.1145/3472810.

[4] J. Lu, Y.-P. Tan, G. Wang, and G. Yang, ''Image-to-set face recognition using locality repulsion projections and sparse reconstruction-based similarity measure,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 6, pp. 1070–1080, Jun. 2013, doi: 10.1109/TCSVT.2013.2241353.

[5] G.-S. Hsu, H.-C. Shie, C.-H. Hsieh, and J.-S. Chan, ''Fast landmark localization with 3D component reconstruction and CNN for cross-pose recognition,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 11, pp. 3194–3207, Nov. 2018, doi: 10.1109/TCSVT.2017.2748379.

[6] J. Lou, X. Cai, J. Dong, and H. Yu, ''Real-time 3D facial tracking via cascaded compositional learning,'' *IEEE Trans. Image Process.*, vol. 30, pp. 3844–3857, Sep. 2021, doi: 10.1109/TIP.2021.3065819.

[7] H. Samani, C.-Y. Yang, C. Li, C.-L. Chung, and S. Li, ''Anomaly detection with vision-based deep learning for epidemic prevention and control,'' *J. Comput. Des. Eng.*, vol. 9, no. 1, pp. 187–200, Feb. 2022.

[8] C. Cao, Q. Hou, and K. Zhou, ''Displaced dynamic expression regression for real-time facial tracking and animation,'' *ACM Trans. Graph.*, vol. 33, no. 4, pp. 1–10, Jul. 2014, doi: 10.1145/2601097.2601204.

[9] C. Cao, H. Wu, Y. Weng, T. Shao, and K. Zhou, ''Real-time facial animation with image-based dynamic avatars,'' *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–12, Jul. 2016, doi: 10.1145/2897824.2925873.

[10] D. S. Alexiadis, A. Chatzitofis, N. Zioulis, O. Zoidi, G. Louizis, D. Zarpalas, and P. Daras, ''An integrated platform for live 3D human reconstruction and motion capturing,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 4, pp. 798–813, Apr. 2017, doi: 10.1109/TCSVT.2016.2576922.

[11] H. Zhu, Y. Liu, J. Fan, Q. Dai, and X. Cao, ''Video-based outdoor human reconstruction,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 4, pp. 760–770, Apr. 2017, doi: 10.1109/TCSVT.2016.2596118.

[12] Z. Zheng, T. Yu, Y. Liu, and Q. Dai, ''PaMIR: Parametric model-conditioned implicit representation for image-based human reconstruction,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3170–3184, Jun. 2022, doi: 10.1109/TPAMI.2021.3050505.

[13] K. Lin, L. Wang, and Z. Liu, ''End-to-end human pose and mesh reconstruction with transformers,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jun. 2021, pp. 1954–1963, doi: 10.1109/CVPR46437.2021.00199.

[14] Y.-W. Zhang, C. Zhang, W. Wang, Y. Chen, Z. Ji, and H. Liu, ''Portrait relief modeling from a single image,'' *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 8, pp. 2659–2670, Aug. 2020, doi: 10.1109/TVCG.2019.2892439.

[15] X. Fan, A. R. Shahid, and H. Yan, ''Facial micro-expression generation based on deep motion retargeting and transfer learning,'' in *Proc. 29th ACM Int. Conf. Multimedia*, New York, NY, USA, Oct. 2021, pp. 4735–4739, doi: 10.1145/3474085.3479210.

[16] J. P. Lewis, K. Anjyo, T. Rhee, M. Zhang, F. H. Pighin, and Z. Deng, ''Practice and theory of blendshape facial models,'' in *Eurographics*. Geneva, Switzerland: The Eurographics Association, 2014, pp. 199–218, doi: 10.2312/egst.20141042.

[17] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, ''FaceWarehouse: A 3D facial expression database for visual computing,'' *IEEE Trans. Vis. Comput. Graphics*, vol. 20, no. 3, pp. 413–425, Mar. 2014, doi: 10.1109/TVCG.2013.249.

[18] H. Li, J. Yu, Y. Ye, and C. Bregler, ''Realtime facial animation with on-the-fly correctives,'' *ACM Trans. Graph.*, vol. 32, no. 4, pp. 1–10, Jul. 2013, doi: 10.1145/2461912.2462019.

[19] L. Sheng, J. Cai, T.-J. Cham, V. Pavlovic, and K. N. Ngan, ''A generative model for depth-based robust 3D facial pose tracking,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jul. 2017, pp. 4598–4607, doi: 10.1109/CVPR.2017.489.

[20] P. Ekman, *Facial Action Coding System (FACS)*. Salt Lake City, UT, USA: A Human Face, 2002.

[21] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, ''The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression,'' in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops*, New York, NY, USA, Jun. 2010, pp. 94–101, doi: 10.1109/CVPRW.2010.5543262.

[22] Z. Abrams and J. Liu, ''Greedy is good: On service tree placement for in-network stream processing,'' in *Proc. 26th IEEE Int. Conf. Distrib. Comput. Syst.*, New York, NY, USA, Jul. 2006, pp. 72–81, doi: 10.1109/ICDCS.2006.45.

[23] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, ''Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit,'' *Signal Process.*, vol. 86, no. 3, pp. 572–588, 2006, doi: 10.1016/j.sigpro.2005.05.030.

[24] G. Brassard and P. Bratley, *Fundamentals of Algorithmics*, vol. 524, 1st ed. Upper Saddle River, NJ, USA: Prentice-Hall, 1996.

[25] P. Civicioglu, ''Backtracking search optimization algorithm for numerical optimization problems,'' *Appl. Math. Comput.*, vol. 219, no. 15, pp. 8121–8144, Apr. 2013, doi: 10.1016/j.amc.2013.02.017.

[26] A. E. Ichim, S. Bouaziz, and M. Pauly, ''Dynamic 3D avatar creation from hand-held video input,'' *ACM Trans. Graph.*, vol. 34, no. 4, pp. 1–14, Jul. 2015, doi: 10.1145/2766974.

[27] S. Bouaziz, Y. Wang, and M. Pauly, ''Online modeling for realtime facial animation,'' *ACM Trans. Graph.*, vol. 32, no. 4, pp. 1–10, Jul. 2013, doi: 10.1145/2461912.2461976.

[28] P. Garrido, M. Zollhöfer, D. Casas, L. Valgaerts, K. Varanasi, P. Pérez, and C. Theobalt, ''Reconstruction of personalized 3D face rigs from monocular video,'' *ACM Trans. Graph.*, vol. 35, no. 3, pp. 1–15, Jun. 2016, doi: 10.1145/2890493.

[29] S. Kim, S. Jung, K. Seo, R. B. I. Ribera, and J. Noh, ''Deep learning-based unsupervised human facial retargeting,'' *Comput. Graph. Forum*, vol. 40, no. 7, pp. 45–55, Nov. 2021, doi: 10.1111/cgf.14400.

[30] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, and S. Zafeiriou, ''3D face morphable models 'in-the-wild,''' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jul. 2017, pp. 5464–5473, doi: 10.1109/CVPR.2017.580.

[31] M. Piotraschke and V. Blanz, ''Automated 3D face reconstruction from multiple images using quality measures,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jun. 2016, pp. 3418–3427, doi: 10.1109/CVPR.2016.372.

[32] D. Thomas and R.-I. Taniguchi, ''Augmented blendshapes for real-time simultaneous 3D head modeling and facial motion capture,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jun. 2016, pp. 3299–3308, doi: 10.1109/CVPR.2016.359.

[33] L. Tran, F. Liu, and X. Liu, ''Towards high-fidelity nonlinear 3D face morphable model,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jun. 2019, pp. 1126–1135, doi: 10.1109/CVPR.2019.00122.

[34] J. Zhang, K. Chen, and J. Zheng, ''Facial expression retargeting from human to avatar made easy,'' *IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 2, pp. 1274–1287, Feb. 2022, doi: 10.1109/TVCG.2020.3013876.

[35] K. Anjyo, H. Todo, and J. P. Lewis, ''A practical approach to direct manipulation blendshapes,'' *J. Graph. Tools*, vol. 16, no. 3, pp. 160–176, Aug. 2012, doi: 10.1080/2165347X.2012.689747.

[36] C. Cao, Y. Weng, S. Lin, and K. Zhou, ''3D shape regression for real-time facial animation,'' *ACM Trans. Graph.*, vol. 32, no. 4, pp. 1–10, Jul. 2013, doi: 10.1145/2461912.2462012.

[37] H. Li, T. Weise, and M. Pauly, ''Example-based facial rigging,'' *ACM Trans. Graph.*, vol. 29, no. 4, pp. 1–6, Jul. 2010, doi: 10.1145/1778765.1778769.

[38] Y. Seol, J. P. Lewis, J. Seo, B. Choi, K. Anjyo, and J. Noh, ''Spacetime expression cloning for blendshapes,'' *ACM Trans. Graph.*, vol. 31, no. 2, pp. 1–12, Apr. 2012, doi: 10.1145/2159516.2159519.

[39] T. Weise, S. Bouaziz, H. Li, and M. Pauly, ''Realtime performance-based facial animation,'' *ACM Trans. Graph.*, vol. 30, no. 4, pp. 1–10, Jul. 2011, doi: 10.1145/2010324.1964972.

[40] V. Blanz and T. Vetter, ''A morphable model for the synthesis of 3D faces,'' in *Proc. 26th Annu. Conf. Comput. Graph. Interact. Techn.*, New York, NY, USA, 1999, pp. 187–194, doi: 10.1145/311535.311556.

[41] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman, ''Unsupervised training for 3D morphable model regression,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New York, NY, USA, Jun. 2018, pp. 8377–8386, doi: 10.1109/CVPR.2018.00874.

[42] A. T. Tran, T. Hassner, I. Masi, and G. Medioni, ''Regressing robust and discriminative 3D morphable models with a very deep neural network,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jul. 2017, pp. 1493–1502, doi: 10.1109/CVPR.2017.163.

[43] H. Dai, N. Pears, W. Smith, and C. Duncan, "A 3D morphable model of craniofacial shape and texture variation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, New York, NY, USA, Oct. 2017, pp. 3104–3112, doi: 10.1109/ICCV.2017.335.

[44] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou, "GANFIT: Generative adversarial network fitting for high fidelity 3D face reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jun. 2019, pp. 1155–1164, doi: 10.1109/CVPR.2019.00125.

[45] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, New York, NY, USA, Jun. 2019, pp. 1–11, doi: 10.1109/CVPRW.2019.00038.

[46] Z. Bai, Z. Cui, X. Liu, and P. Tan, "Riggable 3D face reconstruction via in-network optimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jun. 2021, pp. 6216–6225, doi: 10.1109/CVPR46437.2021.00615.

[47] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao, "FaceScape: A large-scale high quality 3D face dataset and detailed riggable 3D face prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jun. 2020, pp. 601–610, doi: 10.1109/CVPR42600.2020.00068.

[48] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3D face model from in-the-wild images," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–13, Aug. 2021, doi: 10.1145/3450626.3459936.

[49] 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. 1–17, Nov. 2017, doi: 10.1145/3130800.3130813.

[50] X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li, "Face alignment across large poses: A 3D solution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jun. 2016, pp. 146–155, doi: 10.1109/CVPR.2016.23.

[51] E. Richardson, M. Sela, R. Or-El, and R. Kimmel, "Learning detailed face reconstruction from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jul. 2017, pp. 5553–5562, doi: 10.1109/CVPR.2017.589.

[52] E. Richardson, M. Sela, and R. Kimmel, "3D face reconstruction by learning from synthetic data," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, New York, NY, USA, Oct. 2016, pp. 460–469, doi: 10.1109/3DV.2016.56.

[53] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jun. 2016, pp. 2387–2395, doi: 10.1109/CVPR.2016.262.

[54] F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni, "ExpNet: Landmark-free, deep, 3D facial expressions," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, New York, NY, USA, May 2018, pp. 122–129, doi: 10.1109/FG.2018.00027.

[55] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3D dense face alignment," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2020, pp. 152–168.

[56] J. P. Lewis, Z. Mo, K. Anjyo, and T. Rhee, "Probable and improbable faces," in *Mathematical Progress in Expressive Image Synthesis I*, K. Anjyo, Ed. Tokyo, Japan: Springer, 2014, pp. 21–30.

[57] K. Hyeongwoo, G. Pablo, T. Ayush, X. Weipeng, T. Justus, N. Matthias, P. Patrick, R. Christian, Z. Michael, and T. Christian, "Deep video portraits," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–14, Jul. 2018, doi: 10.1145/3197517.3201283.

[58] M. Luthi, T. Gerig, C. Jud, and T. Vetter, "Gaussian process morphable models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 8, pp. 1860–1873, Aug. 2018, doi: 10.1109/TPAMI.2017.2739743.

[59] P. Koppen, Z.-H. Feng, J. Kittler, M. Awais, W. Christmas, X.-J. Wu, and H.-F. Yin, "Gaussian mixture 3D morphable face model," *Pattern Recognit.*, vol. 74, pp. 617–628, Feb. 2018, doi: 10.1016/j.patcog.2017.09.006.

[60] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt, "Real-time expression transfer for facial reenactment," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–14, Nov. 2015, doi: 10.1145/2816795.2818056.

[61] Y. Liu, F. Xu, J. Chai, X. Tong, L. Wang, and Q. Huo, "Video-audio driven real-time facial animation," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 1–10, Nov. 2015, doi: 10.1145/2816795.2818122.

[62] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3D face reconstruction and dense alignment with position map regression network," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2018, pp. 557–574.

[63] B. Chaudhuri, N. Vesdapunt, L. Shapiro, and B. Wang, "Personalized face modeling for improved face reconstruction and motion retargeting," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2020, pp. 142–160.

[64] J. Kang and S. Lee, "A greedy pursuit approach for fitting 3D facial expression models," *IEEE Access*, vol. 8, pp. 192682–192692, 2020.

[65] L. Ke, X. Li, Y. Bisk, A. Holtzman, Z. Gan, J. Liu, J. Gao, Y. Choi, and S. Srinivasa, "Tactical rewind: Self-correction via backtracking in vision-and-language navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jun. 2019, pp. 6741–6749, doi: 10.1109/CVPR.2019.00690.

[66] Z. Zhang, "Iterative point matching for registration of free-form curves and surfaces," *Int. J. Comput. Vis.*, vol. 13, no. 2, pp. 119–152, 1994, doi: 10.1007/BF01427149.

[67] G. Su, J. Jin, Y. Gu, and J. Wang, "Performance analysis of l_0 norm constraint least mean square algorithm," *IEEE Trans. Signal Process.*, vol. 60, no. 5, pp. 2223–2235, Mar. 2012, doi: 10.1109/TSP.2012.2184537.

[68] S. Sanyal, T. Bolkart, H. Feng, and M. J. Black, "Learning to regress 3D face shape and expression from an image without 3D supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New York, NY, USA, Jun. 2019, pp. 7763–7772, doi: 10.1109/CVPR.2019.00795.

[69] B. Chu, S. Romdhani, and L. Chen, "3D-aided face recognition robust to expression and pose variations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, New York, NY, USA, Jun. 2014, pp. 1899–1906, doi: 10.1109/CVPR.2014.245.

[70] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, "The Mahalanobis distance," *Chemometrics Intell. Lab. Syst.*, vol. 50, no. 1, pp. 1–18, 2000, doi: 10.1016/S0169-7439(99)00047-7.

[71] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM J. Sci. Comput.*, vol. 16, no. 5, pp. 1190–1208, Sep. 1995, doi: 10.1137/0916069.

[72] R. A. Waltz, J. L. Morales, J. Nocedal, and D. Orban, "An interior algorithm for nonlinear optimization that combines line search and trust region steps," *Math. Program.*, vol. 107, no. 3, pp. 391–408, 2006, doi: 10.1007/s10107-004-0560-5.

[73] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3D face analysis," in *Proc. Eur. Workshop Biometrics Identity Manage.* Berlin, Germany: Springer, 2008, pp. 47–56.

[74] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affect. Comput.*, vol. 10, no. 1, pp. 18–31, Jan. 2019, doi: 10.1109/TAFFC.2017.2740923.

[75] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *Proc. 6th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, New York, NY, USA, Sep. 2009, pp. 296–301, doi: 10.1109/AVSS.2009.58.

[76] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, New York, NY, USA, Jun. 2013, pp. 532–539, doi: 10.1109/CVPR.2013.75.

[77] P. Huber, Z.-H. Feng, W. Christmas, J. Kittler, and M. Ratsch, "Fitting 3D morphable face models using local features," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, New York, NY, USA, Sep. 2015, pp. 1195–1199, doi: 10.1109/ICIP.2015.7350989.

[78] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: Database and results," *Image Vis. Comput.*, vol. 47, pp. 3–18, Mar. 2016, doi: 10.1016/j.imavis.2016.01.002.

[79] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 FPS via regressing local binary features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, New York, NY, USA, Jun. 2014, pp. 1685–1692, doi: 10.1109/CVPR.2014.218.

[80] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn, "DISFA: A spontaneous facial action intensity database," *IEEE Trans. Affect. Comput.*, vol. 4, no. 2, pp. 151–160, Apr. 2013, doi: 10.1109/T-AFFC.2013.4.

[81] A. Mehrabian and J. A. Russell, *An Approach to Environmental Psychology*. Cambridge, MA, USA: MIT Press, 1974.

[82] D. Bradley, W. Heidrich, T. Popa, and A. Sheffer, "High resolution passive facial performance capture," *ACM Trans. Graph.*, vol. 29, no. 4, pp. 1–10, Jul. 2010, doi: 10.1145/1778765.1778778.

[83] T. Beeler, F. Hahn, D. Bradley, B. Bickel, P. Beardsley, C. Gotsman, R. W. Sumner, and M. Gross, "High-quality passive facial performance capture using anchor frames," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 1–10, Jul. 2011, doi: 10.1145/2010324.1964970.

[84] Z. Sun, Y. He, A. Gritsenko, A. Lendasse, and S. Baek, "Embedded spectral descriptors: Learning the point-wise correspondence metric via Siamese neural networks," *J. Comput. Des. Eng.*, vol. 7, no. 1, pp. 18–29, Feb. 2020.

JIWOO KANG received the B.S. degree in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2011, and the M.S. and Ph.D. degrees in electrical and electronic engineering from Yonsei University, in 2019, through the integrated Ph.D. program. He worked as a Researcher at Yonsei University, from September 2019 to November 2020, where he was a Research Professor at the Y-BASE R&E Institute, from December 2020 to February 2022. He is currently an Assistant Professor with Sookmyung Women's University, Seoul. His research interests include computer graphics, computer vision, and image processing.

HYEWON SONG received the B.S. degree from Hongik University, Seoul, South Korea, in 2015. She is currently pursuing the Ph.D. degree with the Department of Electrical and Electronic Engineering, Yonsei University, Seoul. Her research interests include areas of computer vision and machine learning.

KYOUNGOH LEE received the B.S. degree in electronic engineering from Soongsil University, Seoul, South Korea, in 2014, and the M.S. and Ph.D. degrees from the Multidimensional Insight Laboratory, Yonsei University, Seoul, in 2021. He was a Research Assistant under the guidance of Prof. Kot Chichung Alex with the Laboratory for School of Computer Science and Engineering, Nanyang Technological University (NTU), Singapore, in 2018. Currently, he is with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea. His research interests include deep learning, computer vision, human pose estimation, human behavior understanding, and video surveillance.

SANGHOON LEE (Senior Member, IEEE) received the B.S. degree from Yonsei University, South Korea, in 1989, the M.S. degree from KAIST, South Korea, in 1991, and the Ph.D. degree from The University of Texas at Austin, TX, USA, in 2000. From 1991 to 1996, he was with Korea Telecom, South Korea. From 1999 to 2002, he was with Lucent Technologies, NJ, USA. In 2003, he joined the Department of Electrical Engineering (EE), Yonsei University, as a Faculty Member, where he is currently a Full Professor. His current research interests include image/video processing, computer vision, and graphics. He was a member of the IEEE IVMSP/MMSP TC, from 2014 to 2019 and from 2016 to 2021. He is also a BoG Member of APSIPA and the Editor-in-Chief of APSIPA News Letters. He was the General Chair of the 2013 IEEE IVMSP Workshop. He has been serving as the Chair for the IEEE P3333.1 Working Group, since 2011. He was the Image, Video, and Multimedia TC Chair of APSIPA, from 2018 to 2019. He served as an Editor for the *Journal of Communications and Networks*, from 2009 to 2015. He was an Associate Editor and a Guest Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, from 2010 to 2014 and in 2013, respectively. He served as an Associate Editor and has been serving as a Senior Area Editor for the IEEE SIGNAL PROCESSING LETTERS, from 2014 to 2018 and since 2018, respectively. He has been serving as an Associate Editor of IEEE TRANSACTIONS ON MULTIMEDIA, since 2022.