Generating Multiple 4D Expression Transitions by Learning Face Landmark Trajectories

Generating Multiple 4D Expression Transitions by Learning Face Landmark Trajectories

Naima Otberdout*, Claudio Ferrari*, Mohamed Daoudi, Stefano Berretti, Alberto Del Bimbo,

Abstract—In this work, we address the problem of 4D facial expressions generation. This is usually addressed by animating a neutral 3D face to reach an expression peak, and then get back to the neutral state. In the real world though, people show more complex expressions, and switch from one expression to another. We thus propose a new model that generates transitions between different expressions, and synthesizes long and composed 4D expressions. This involves three sub-problems: *(i)* modeling the temporal dynamics of expressions, *(ii)* learning transitions between them, and *(iii)* deforming a generic mesh. We propose to encode the temporal evolution of expressions using the motion of a set of 3D landmarks, that we learn to generate by training a manifold-valued GAN (Motion3DGAN). To allow the generation of composed expressions, this model accepts two labels encoding the starting and the ending expressions. The final sequence of meshes is generated by a Sparse2Dense mesh Decoder (S2D-Dec) that maps the landmark displacements to a dense, per-vertex displacement of a known mesh topology. By explicitly working with motion trajectories, the model is totally independent from the identity. Extensive experiments on five public datasets show that our proposed approach brings significant improvements with respect to previous solutions, while retaining good generalization to unseen data.

Index Terms—4D Facial Expression generation, facial landmarks, 3D meshes.

1 Introduction

Generating dynamic 3D (4D) face models is the task of synthesizing realistic 3D face instances that dynamically evolve across time with varying expressions or speech-related movements, while keeping the same identity. This can be useful in a wide range of graphics applications, spanning from 3D face modeling to augmented and virtual reality for animated films and computer games. While recent advances in generative neural networks have made possible the development of effective solutions that operate on 2D images [1], [2], the literature on the problem of generating facial animation in 3D is still quite limited, with few examples available [3], [4]

Performing faithful and accurate 3D facial animations requires addressing some major challenges, in terms both of 3D face modeling, and temporal dynamics. Related to the former, as we wish to animate a 3D face of an individual, its identity should be maintained across time. Also, the applied dynamic deformation should be controllable, corresponding to a specific expression/motion, and should be applicable to any 3D face. Incidentally, these are major challenges in 3D face modeling, which require disentangling structural face elements related to the identity, *e.g.*, nose or jaw shape, from deformations related to the movable face parts, *e.g.*, mouth opening/closing. Modeling the temporal dynamics of expressions, instead, gives rise to other challenges. First, expressions are very personal, and different individuals might perform the same expressions differently; for example, it is difficult to find two people laughing in the same exact way. So, for the sake of realism, one should be able to generate diverse motions even for the same expression. Furthermore, dynamic expressions have been always standardized in a way that they are supposed to start from a neutral configuration, reach (onset) a peak of the expression (apex), and then get back (offset) to neutral [5], [6], [7]. However, this is not the case in the real world, where people might switch from one expression to another dynamically, so that a sequence could start from or end to a non-neutral configuration. This implies considering the sequence as a transition between two expressions.

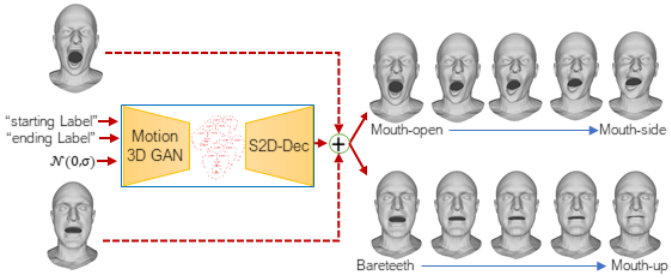


Fig. 1. **3D dynamic facial expression generation**: A GAN generates the motion of 3D landmarks from a pair of expression labels, *i.e.*, one starting and one ending labels and noise; A decoder expands the animation from the landmarks to a dense mesh, while keeping the identity of a neutral 3D face.

Some previous works tackled the problem of neutral-to-apex generation by capturing the facial expression of a subject frame-by-frame and transferring it to a target model [8]. However, in this

- * *Equal contributions.*
- *N. Otberdout is with Ai movement - University Mohammed VI Polytechnic, Rabat, Morocco, E-mail: naima.otberdout@um6p.ma*
- *C. Ferrari is with the Department of Architecture and Engineering University of Parma, Italy, E-mail: claudio.ferrari2@unipr.it*
- *M. Daoudi is with Univ. Lille, CNRS, Centrale Lille, Institut Mines-Télécom, UMR 9189 CRIStAL, F-59000 Lille, France and IMT Nord Europe, Institut Mines-Télécom, Univ. Lille, Centre for Digital Systems, F-59000 Lille, France, E-mail: mohamed.daoudi@imt-nord-europe.fr*
- *S. Berretti and A. Del Bimbo are with Media Integration and Communication Center (MICC), Univ. of Florence, Italy. E-mail: {stefano.berretti, alberto.delbimbo}@unifi.it*

Manuscript received July 14, 2022; revised ????

case, the temporal evolution is neglected, so the problem reduces to transferring a tracked expression to a neutral 3D face. Some other works animated a 3D face mesh given an arbitrary speech signal and a static 3D face mesh as input [9], [10], some also considering additional emotional labels [11], [12], [13]. Also in this case though, the temporal evolution is guided by an external input, similar to a tracked expression. Differently, here we are interested in animating a face just starting from a generic 3D face and a pair of expression labels, *i.e.*, a starting and an ending label.

In our solution, which is illustrated in Figure 1, the temporal evolution and the mesh deformation are decoupled and modeled separately in two network architectures. A manifold-valued GAN (*Motion3DGAN*) accounts for the expression dynamics by generating a temporally consistent motion of 3D landmarks corresponding to a specified transition between two expressions from a noise signal. The landmarks motion is encoded using the *Square Root Velocity Function* (SRVF), and compactly represented as a point on a hypersphere. The novel characteristic of this network is that, differently from all the previous literature, it can generate motions that do not necessarily start from a neutral configuration. In particular, the sequence to be modeled is specified by two labels, encoding the starting face configuration, and the ending one. In this way, we are able to generate *(i)* arbitrarily long, and *(ii)* composed motions of landmarks. Then, a Sparse2Dense mesh Decoder (*S2D-Dec*) generates a dense 3D face guided by the landmarks motion for each frame of the sequence.Ultimately, the two networks allow us to generate a dynamic sequence of 3D faces performing a dynamic transitions between expressions. To effectively disentangle identity and expression components, the landmarks motion is represented as a per-frame displacement (motion) from the starting configuration. Instead of directly generating a mesh, the S2D-Dec expands the landmarks displacement to a dense, per-vertex displacement, which is finally used to deform the neutral mesh. We thus train the decoder to learn how the displacement of a sparse set of points influences the displacement of the whole face surface. This has the advantage that structural face parts, *e.g.*, nose or forehead, which are not influenced by facial expressions are ignored, helping in maintaining the identity traits stable. Furthermore, the network can focus on learning expressions at a fine-grained level of detail, and generalize to unseen identities.

In summary, the main contributions of our work are: *(i)* we propose an original method to generate dynamic sequences of 3D expressive scans given a 3D face mesh and a pair of expression labels representing, respectively, the starting and ending expression of the sequence. This is the first solution that can generate smooth transitions in 3D between two generic expression labels, while works in the literature are constrained to the neutral-to-apex transition. Our approach can generate strong and diverse expression sequences, with high generalization ability to unseen identities and expressions. This has been obtained by adapting the GAN architecture proposed in [4] for accepting and learning from two labels. Doing so demanded for a dataset including transitions between expressions. Given that such dataset does not exist, we *(ii)* defined a data augmentation strategy specific for 4D expression sequences, which is based on the SRVF encoding; finally, *(iii)* we exploit the above characteristic of our model to generate concatenated sequences of expression transitions. An overall sequence can start from an expression A, change to an expression B, then to expressions C and D. This is modeled as a combined generation from the pair $A-B$ to $B-C$ and $C-D$. We prove the expressions generated in the subsequent stages of the generation, *i.e.*, $B-C$ and $C-D$, do not diverge though they are generated starting from the synthetic model generated at the end of the previous transition.

The rest of the paper is organized as follows: In Section 2, we summarize the works in the literature that are closer to our proposed solution; In Section 3, we introduce the proposed solution for generating the temporal dynamics of facial landmarks and to derive a dense mesh from them; A comprehensive experimental evaluation of our approach is presented in Section 4; Finally, conclusions and future work directions are given in Section 5.

2 Related Work

Our work is related to methods for *(a)* 3D face modeling, *(b)* facial expression generation guided by landmarks, and *(c)* dynamic generation of 3D faces, *i.e.*, 4D face generation. Below, we summarize works in these three areas that are relevant for our proposal.

3D face modeling

The 3D Morphable face Model (3DMM) as originally proposed in [14] is the most popular solution for modeling 3D faces. The original model and its variants [15], [16], [17], [18], [19], [20], [21] capture face shape variations both for identity and expression based on linear formulations, thus incurring in limited modeling capabilities. For this reason, non-linear encoder-decoder architectures are attracting more and more attention. This comes at the cost of reformulating convolution and pooling/unpooling like operations on the irregular mesh support [22], [23], [24]. For example, Ranjan *et al.* [7] proposed an auto-encoder architecture that builds upon newly defined spectral convolution operators, and pooling operations to down-/up-sample the mesh. Bouritsas *et al.* [25] improved upon the above by proposing a graph convolutional operator enforcing consistent local orderings on the vertices of the graph through the *spiral operator* [26]. Despite their impressive modeling precision, a recent work [20] showed that they heavily suffer from poor generalization to unseen identities. This limits their practical use in tasks such as face fitting or expression transfer. We finally mention that other approaches do exist to learn generative 3D face models, such as [27], [28]. However, instead of dealing with meshes they use alternative representations for 3D data, such as depth images or UV-maps.

To overcome the above limitation, we go beyond self-reconstruction and propose a mesh decoder that, differently from previous models, learns expression-specific mesh deformations from a sparse set of landmark displacements.

Facial expression generation guided by landmarks

Recent advances in neural networks made facial landmark detection reliable and accurate both in 2D [29], [30], [31] and 3D [32], [33]. Landmarks and their motion are a viable way to account for facial deformations as they reduce the complexity of the visual data, and have been commonly used in several 3D face related tasks, *e.g.*, reconstruction [21], [34] or reenactment [35], [36]. Despite some effort was put in developing landmark-free solutions for 3D face modeling [37], [38], [39], some recent works investigated their use to model the dynamics of expressions. Wang *et al.* [40] proposed a framework that decouples facial expression dynamics, encoded into landmarks, and face appearance using a

conditional recurrent network. Otberdout *et al.* [2] proposed an approach for generating videos of the six basic expressions given a neutral face image. The geometry is captured by modeling the motion of landmarks with a GAN that learns the distribution of expression dynamics.

These methods demonstrated the potential of using landmarks to model the dynamics of expressions and generate 2D videos. In our work, we instead tackle the problem of modeling the dynamics in 3D, exploring the use of 3D landmarks motion to both model the temporal evolution of expressions and animate a 3D face.

4D face generation

While many researchers tackled the problem of 3D mesh deformation, the task of 3D facial motion synthesis is yet more challenging. A few studies addressed this issue by exploiting audio features [10], [41], speech signals [9] or tracked facial expressions [8] to generate facial motions. However, none of these explicitly models the temporal dynamics.

The work in [3] first addressed the problem of dynamic 3D expression generation. In that framework, the motion dynamics is modeled with a temporal encoder based on an LSTM, which produces a per-frame latent code starting from a per-frame expression label. The codes are then fed to a mesh decoder that, similarly to our approach, generates a per-vertex displacement that is summed to a neutral 3D face to obtain the expressive meshes. Despite the promising results reported in [3], we identified some limitations in this solution. First, the LSTM is deterministic, and for a given label the exact same displacements are generated. Our solution instead achieves diversity in the output sequences by generating from noise. Moreover, in [3] the mesh decoder generates the displacements from the latent codes, making it dependent from the temporal encoder. In our solution, the motion dynamics and mesh displacement generation are decoupled, using landmarks to link the two modules. The S2D-Dec is thus independent from Motion3DGAN, and can be used to generate static meshes as well given an arbitrary set of 3D landmarks as input. This permits us to use the decoder for other tasks such as expression/speech transfer. Finally, as pointed out in [3], the model cannot perform extreme variations well. Using landmarks allowed us to define a novel reconstruction loss that weighs the error of each vertex with respect to its distance from the landmarks, encouraging accurate modeling of the movable parts. Thanks to this, we are capable of accurately reproducing from slight to strong expressions, and generalize to unseen motions.

This work develops on the generative model proposed in Otberdout *et al.* [4]. Compared to this previous approach, the main novelties of this paper are:

- we removed the constraint of starting the 4D sequence from a neutral face. Motion3DGAN was modified so that it can generate 4D transitions that switch between two generic expressions;
- we defined a strategy to augment the dataset of 4D expressions with interpolated, complex expressions;
- we expanded the experimental validation to three additional datasets, characterized by totally different expressions, identities and mesh topology;
- we experimented more difficult scenarios, such as speech transfer and cross-dataset 3D reconstruction.

3 Proposed Method

Our approach consists of two specialized networks as summarized in Figure 2. Motion3DGAN accounts for the temporal dynamics and generates the motion of a sparse set of 3D landmarks from noise. The generated motion represents a transition between two expressions defined by two labels, one for the start, *e.g.*, neutral, happy, and the other for the ending configuration. The motion is then converted as a per-frame landmarks displacement. These displacements are then fed to a decoder network (S2D-Dec) that constructs the dense point-cloud displacements from the sparse displacements given by the landmarks. These dense displacements are finally added to a generic 3D face to generate a sequence of 3D faces corresponding to the specified transition from the starting expression to the ending one. In the following, we separately describe the two networks.

3.1 Generating 4D Expressions: Motion3DGAN

Facial landmarks were shown to well encode the temporal evolution of facial expressions [2], [42]. Motivated by this fact, we generate the facial expression dynamics based on the motion of 3D facial landmarks. Given a set of k 3D landmarks, $Z(t) = (x_i(t), y_i(t), z_i(t))_{i=1}^{k}$, with $Z(0)$ being the starting configuration, their motion can be seen as a trajectory in $\mathbb{R}^{k\times 3}$, and can be formulated as a parameterized curve in $\mathbb{R}^{k\times 3}$ space. Let $\alpha : I = [0, 1] \rightarrow \mathbb{R}^{k\times 3}$ represent the parameterized curve, where each $\alpha(t) \in \mathbb{R}^{k\times 3}$. For the purpose of modeling and studying our curves, we adopt the Square-Root Velocity Function (SRVF) proposed in [43]. The SRVF $q(t) : I \rightarrow \mathbb{R}^{k\times 3}$ is defined by:

$$q(t) = \begin{cases} \frac{\dot{\alpha}(t)}{\sqrt{\|\dot{\alpha}(t)\|}}, & \text{if } \|\dot{\alpha}(t)\| \neq 0 \\ 0, & \text{if } \|\dot{\alpha}(t)\| = 0, \end{cases} \tag{1}$$

where, $\|\cdot\|$ is the Euclidean 2-norm in $\mathbb{R}^n$. This function proved effective for tasks such as human action recognition [44] or 3D face recognition [45]. Similar to this work, Otberdout *et al.* [2] proposed to use the SRVF representation to model the temporal evolution of 2D facial landmarks, which makes it possible to learn the distribution of these points and generate the motions for new 2D facial expression. In this paper, we extend this idea by proposing the Motion3DGAN model, which generates the motion of 3D facial landmarks. Differently from [2], where the dynamic expression is assumed to start from a neutral configuration, here we remove this constraint and train Motion3DGAN to generate motions corresponding to a transition between two expressions. The motion is represented using the SRVF encoding in (1). Following [2], we remove the scale variability of the resulting motions by restricting curves α to length 1. As a result, we transform the motion of 3D facial landmarks to points on a Hilbert sphere of radius 1, $\mathcal{C} = \{q : [0, 1] \rightarrow \mathbb{R}^{k\times 3}, \|q\|^2 = 1\}$. The geometry of sphere is well-understood and can be exploited.

To learn the distribution of the SRVF representations, we propose Motion3DGAN as an extension of MotionGAN [2], a conditional version of the Wasserstein GAN for manifold-valued data [46]. It maps a random vector z to a point on the Hilbert sphere $\mathcal{C}$ conditioned on an input labels pair $c = (start, end)$. Motion3DGAN is composed of two networks trained adversarially: a generator G that learns the distribution of the 3D landmark motions, and a discriminator D that distinguishes between real and generated 3D landmark motions. Motion3DGAN is trained by a

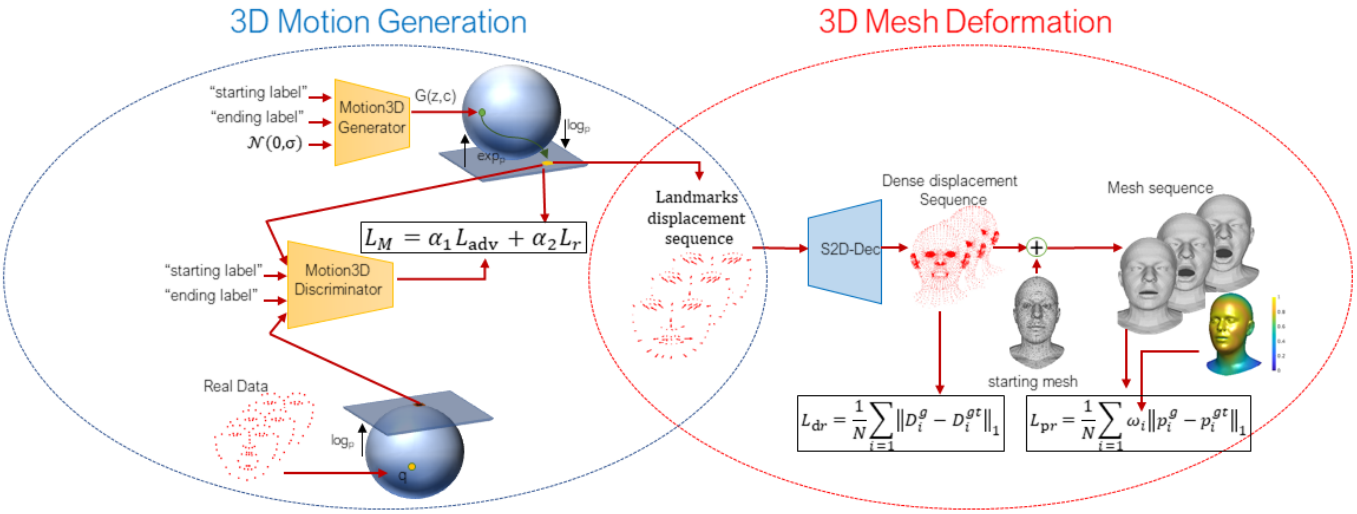


Fig. 2. **Overview of our framework**: Motion3DGAN generates the motion $q(t)$ of 3D landmarks corresponding to an expression label from a noise vector z. The module is trained guided by a reconstruction loss L_r and adversarial loss L_{adv}. The motion $q(t)$ is converted to a sequence of landmark displacements d_i, which are fed to S2D-Dec. From each d_i, the decoder generates a dense displacement D_i^g. A neutral mesh is then summed to the dense displacements to generate the expressive meshes S^g. S2D-Dec is trained under the guidance of a displacement loss L_{dr} and our proposed weighted reconstruction loss L_{pr}.

weighted sum of an adversarial loss L_{adv} and a reconstruction loss L_r such that $L_M = \alpha_1 L_{adv} + \alpha_2 L_r$.

The adversarial loss is L_{adv} formulated as:

$$\begin{aligned} L_{adv} = \; & \mathbb{E}_{q\sim\mathbb{P}_q}\left[D\left(\log_p(q), c\right)\right] \\ & -\mathbb{E}_{z\sim\mathbb{P}_z}\left[D\left(\log_p\left(\exp_p(G(z,c))\right)\right)\right] \\ & +\lambda E_{\hat{q}\sim\mathbb{P}_{\hat{q}}}\left[\left(\|\nabla_{\hat{q}} D(\hat{q})\|_2 - 1\right)^2\right]. \end{aligned} \tag{2}$$

In the above equation, the exponential map, $\exp_u(.)$: $T_u(C) \mapsto \mathcal{C}$ and its inverse, *i.e.*, the logarithm map $\log_u(q)$: $\mathcal{C} \mapsto T_u(C)$ are used to map the SRVF data forth and back to a tangent space T_u defined at a particular point p of $\mathcal{C}$. They are computed as follows:

$$\exp_u(s) = cos(\|s\|)u + sin(\|s\|)\frac{s}{\|s\|}, \tag{3}$$

$$\log_u(q) = \frac{d_{\mathcal{C}}(q,u)}{sin(d_{\mathcal{C}}(q,r))}(q - cos(d_{\mathcal{C}}(q,u))u), \tag{4}$$

where $d_{\mathcal{C}}(q,p) = \cos^{-1}(\langle q,p\rangle)$ is the geodesic distance between q and p in $\mathcal{C}$. In (2), $q \sim \mathbb{P}_q$ is an SRVF sample from the training set, c is the expression labels pair (*e.g.*, mouth open-eyebrow, bareteeth-mouth up) that is concatenated to a random noise $z \sim \mathbb{P}_z$. The last term of the adversarial loss represents the gradient penalty of the Wasserstein GAN [47]. Specifically, $\hat{q} \sim \mathbb{P}_{\hat{q}}$ is a random point sampled uniformly along straight lines between pairs of points sampled from $\mathbb{P}_q$ and the generated distribution $\mathbb{P}_g$:

$$\hat{q} = (1-\tau)\log_p(q) + \tau \log_p(\exp_p(G(z,c))), \tag{5}$$

where $0 \leq \tau \leq 1$, and $\nabla_{\hat{q}} D(\hat{q})$ is the gradient *w.r.t.* $\hat{q}$.

Finally, the reconstruction loss is defined as:

$$L_r = \|\log_p(\exp_p(G(z,c))) - \log_p(q)\|_1, \tag{6}$$

where $\|.\|_1$, represents the L_1-norm, and q is the ground truth SRVF corresponding to the condition c. The generator and discriminator architectures are similar to [2].

The SRVF representation is reversible, which makes it possible to recover the curve $\alpha(t)$ from a new generated SRVF $q(t)$ by,

$$\alpha(t) = \int_0^t \|q(s)\|q(s)ds + \alpha(0), \tag{7}$$

where $\alpha(0)$ represents the initial landmark configuration $Z(0)$. Using this equation, we can apply the generated motion to *any* landmark configuration, making it robust to identity changes.

3.2 From Sparse to Dense 3D Expressions: S2D-Dec

Our final goal is to animate the starting mesh S^n to obtain a novel 3D face S^g reproducing some expression, yet maintaining the identity structure of S^n. Given this, we point at generating the displacements of the mesh vertices from the sparse displacements of the landmarks to animate S^n. In the following, we assume all the meshes have a fixed topology, and are in full point-to-point correspondence.

Let $\mathcal{L} = \left\{\left(S_1^n, S_1^{gt}, Z_1^n, Z_1^{gt}\right), \ldots, \left(S_m^n, S_m^{gt}, Z_m^{gt}, Z_m^n\right)\right\}$ be the training set, where $S_i^n = (p_1^n, \ldots, p_N^n) \in \mathbb{R}^{N\times 3}$ is a neutral 3D face, $S_i^{gt} = (p_1^{gt}, \ldots, p_N^{gt}) \in \mathbb{R}^{N\times 3}$ is a 3D expressive face, $Z_i^n \in \mathbb{R}^{k\times 3}$ and $Z_i^{gt} \in \mathbb{R}^{k\times 3}$ are the 3D landmarks corresponding to S_i^n and S_i^{gt}, respectively. We transform this set to a training set of sparse and dense displacements, $\mathcal{L} = \{(D_1, d_1), \ldots, (D_m, d_m)\}$ such that, $D_i = S_i^{gt} - S_i^n$ and $d_i = Z_i^{gt} - Z_i^n$. Our goal here is to find a mapping $h : \mathbb{R}^{k\times 3} \rightarrow \mathbb{R}^{N\times 3}$ such that $D_i \approx h(d_i)$. We designed the function h as a decoder network (S2D-Dec), where the mapping is between a sparse displacement of a set of landmarks and the dense displacement of the entire mesh points. Finally, in order to obtain the expressive mesh, the dense displacement map is summed to a 3D face in neutral expression, *i.e.*, $S_i^e = S_i^n + D_i$. The S2D-Dec network is based on the spiral operator proposed in [25]. Our architecture includes five spiral convolution layers, each one followed by an up-sampling layer. More details on the architecture can be found in the supplementary material.

In order to train this network, we propose to use two different losses, one acting directly on the displacements and the other

controlling the generated mesh. The reconstruction loss of the dense displacements is given by,

$$L_{dr} = \frac{1}{N}\sum_{i=1}^{N}\left\|D_i^g - D_i^{gt}\right\|_1, \quad (8)$$

where D^g and D^{gt} are the generated and the ground truth dense displacements, respectively. To further improve the reconstruction accuracy, we add a loss that minimizes the error between S^g and the ground truth expressive mesh S^{gt}. We observed that vertices close to the landmarks are subject to stronger deformations. Other regions like the forehead, instead, are relatively stable. To give more importance to those regions, we defined a weighted version of the $L1$ loss:

$$L_{pr} = \frac{1}{N}\sum_{i=1}^{N} w_i \cdot \left\|p_i^g - p_i^{gt}\right\|_1. \quad (9)$$

We defined the weights as the inverse of the Euclidean distance of each vertex p_i in the mesh from its closest landmark Z_j, *i.e.*$w_i = \frac{1}{\min d(p_i, Z_j)}$, $\forall j$. This provides a coarse indication of how much each p_i contributes to the expression generation. Since the mesh topology is fixed, we can pre-compute the weights w_i and re-use them for each sample. Weights are then re-scaled so that they lie in $[0, 1]$. Vertices corresponding to the landmarks, *i.e.*, $p_i = Z_j$ for some j, are hence assigned the maximum weight. We will show this strategy provides a significant improvement with respect to the standard $L1$ loss. The total loss used to train the S2D-Dec is given by $L_{S2D} = \beta_1.L_{dr} + \beta_2.L_{pr}$.

4 Experiments

We validated the proposed method in a broad set of experiments on five publicly available benchmark datasets.

CoMA dataset [7]: It is a common benchmark employed in other studies [7], [25]. It includes 12 subjects, each one performing 12 extreme and asymmetric expressions. Each expression comes as a sequence of meshes $S \in \mathbb{R}^{N\times 3}$ (140 meshes on average), with $N = 5,023$ vertices. Sequences start from a neutral state, reach a peak of the expression, and then get back to a neutral state.

D3DFACS dataset [6]. We used the registered version of this dataset [48], which has the same topology of CoMA. It contains 10 subjects, each one performing a different number of facial expressions. In contrast to CoMA, this dataset is labeled with the activated action units of the performed facial expression. It is worthy to note that the expressions of D3DFACS are highly different from those in CoMA.

Florence 4D Facial Expression dataset (Florence 4D) [49]. This dataset consists of 10,710 synthetic sequences of 3D faces with different facial expressions from which we selected 1,222 sequences corresponding to the 7 standard facial expressions: angry, disgust, fear, happy, sad and surprise. The sequences correspond to 155 subjects including 117 females and 38 males. Each sequence is composed of 60 frames showing an expression that evolves from neutral face to reach the peak and then get back to the neutral state. The meshes are in full correspondence with the Flame template. The dataset includes synthetic identities based on the DAZ Studio's Genesis 8 Female [50] as well as CoMa identities and real scans from the Florence 2D/3D dataset [51]. Expressions were generated with the DAZ Studio software [50].

VOCASET. This dataset provides 480 speech sequences of 3D face scans belonging to the 12 identities of CoMA dataset. The faces are in full correspondence and aligned to the Flame template.

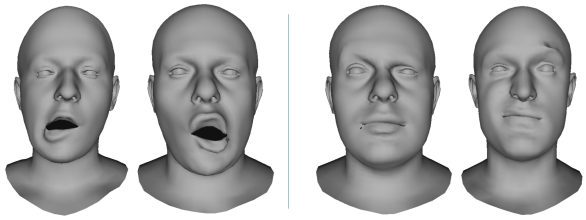


Fig. 3. Samples from the CoMA dataset: Mouth-Side (left), Eyebrows (right). For the same expressions, the samples differ significantly.

BU-3DFE. This dataset contains scans of 44 females and 56 males, ranging from 18 to 70 years old, acquired in neutral plus the prototypical six expressions. Each of the six expressions is acquired at four levels of intensity. Those, however, are not in full, point-to-point correspondence. For the sake of this work, we employed the registered version as described in [21], which includes $1,779$ meshes, each mesh having $N = 6704$ vertices. We underline that we chose this particular dataset in addition to the previous ones to show our S2D-Dec can effectively handle different mesh topology, and is robust to possible noise as can result from a dense registration process.

4.1 Training Details

In order to keep Motion3DGAN and S2D-Dec decoupled, they are trained separately.

Motion3DGAN: We used CoMA to train Motion3DGAN, since this dataset contains 4D sequences labeled with facial expression classes. However, in order to train Motion3DGAN to generate transitions that shift from one expression to another, the CoMA dataset in its original form is not suitable, since it only includes *neutral-peak-neutral* sequences, whereas we also need *peak-peak* transitions. So, we defined a solution to expand the dataset to include such peak-peak transitions.

In particular, we processed the dataset as follows: first, we manually divided the existing sequences into sub-sequences of 30 frames/meshes. They are separated in neutral to peak of the expression (neutral-peak), and vice versa (peak-neutral). These are then encoded as points q on a hyper-sphere using the SRVF representation in (1). Note that, for neutral-peak sequences, the initial landmark configuration required to compute the SRVF motion for a subject j is the neutral one, *i.e.*, $Z_j(0) = Z_j^n$, while for peak-neutral motions, the initial landmark configuration needs to be that of the peak frame of the expression, *i.e.*, $Z_j(0) = Z_j^e$. In this way, we can easily interpolate two sequences: given two points on the sphere q_1 and q_2, representing two expression motions e_1, e_2 (either neutral-peak or peak-neutral), the geodesic path $\psi(\tau)$ between them is given by:

$$\psi(\tau) = \frac{1}{sin(\theta)} sin((1-\tau)\theta)q_1 + sin(\tau\theta)q_2, \quad (10)$$

where, $\theta = d_{\mathcal{C}}(q_1, q_2) = \cos^{-1}(\langle q_1, q_2\rangle)$, and $\tau \in [0, 1]$. This path determines all the points q_i existing between q_1 and q_2, each of them corresponding to a (interpolated) landmarks motion. We do so for all the 12 subjects and their 12 expressions.

We then estimate the transition between two expression peaks. The idea is that each interpolated motion q_i ends to an expression peak, which is a linear combination of the peaks of two generic

Fig. 4. Example of the generation of a peak-peak transition across two given expression sequences. Given two expression motions, q_1 and q_2, we generate 30 interpolated sequences by sampling the geodesic path between their SRVF representation. The transition between the two peaks is obtained by converting back the motion to landmarks, and concatenating the last frame of each sequence (red rectangle). Meshes generated with S2D-Dec are visualized in place of landmarks for clarity.

motions q_1, q_2. Thus, if we generate a certain number of motions q_i at different interpolation steps, convert them back to landmarks using (7), and keep the last frame of each sequence q_i, then their concatenation provides us a transition between the peaks of q_1 and q_2. In order to collect compatible sequences, we generated 30 interpolated motions q_i for each possible pair, so that the peak-peak transition is of the same length of the original ones, *i.e.*, 30 frames. This process is depicted in Figure 4. This sequence is then converted to a point in the SRVF space for training Motion3DGAN. Doing so for all possible pairs resulted in approximately $300K$ samples in total. We chose to perform interpolation directly in the hypersphere induced by SRVF as it leads to cleaner results than interpolating directly in the 3D space [52]. However, we cannot use all of them. In fact, different subjects might perform the same expression in a significantly different way, as shown in Figure 3. To obtain a correct peak-peak transition though, the initial landmark configuration $Z(0)$ must be coherent with the estimated motion, which can only be guaranteed by using identity-specific landmarks. In order to maintain full independence from the identity, we instead computed the average landmark configuration of the expression peaks across subjects, for each expression. We then used these prototypes to select the most similar ones among the interpolated transitions, obtaining a total of $6,740$ interpolated motions. One sample for each transition is used as test set. We used all the others for training as we generate from a random noise at test time.

To train the model, we encoded the motions of $k = 68$ landmarks in the SRVF representation. The landmarks were first centered and normalized to unit norm. To encode the starting-ending labels pair, each of the 13 expression (including neutral) was first encoded as a one-hot vector; the labels pairs is formed by concatenating two expression labels. Finally, they are further concatenated with a random noise vector of size 128.

S2D-Dec: To comprehensively evaluate the capability of S2D-Dec of generalizing to either unseen identities or expressions, we performed subject-independent and expression-independent cross-validation experiments. For the *subject-independent* experiment, we used a 4-fold cross-validation protocol for CoMA, training on 9 and testing on 3 identities in each fold. On D3DFACS, we used the last 7 identities for training and the remaining 3 as test set. For the Florence 4D dataset, we used the last 30 females and 10 males as test set, and trained the model on the remaining subjects. Finally, the test set of the BU-3DFE includes the first 10 subjects (5 females and 5 males). For the *expression-independent* splitting, we used a 4-fold cross-validation protocol for CoMA, training on 9 and testing on 3 expressions in each fold. For D3DFACS, given the different number of expressions per subject, the first 11 expressions were used for testing and the remaining expressions were used for training. For Florence 4D, the first two expressions of each subject were used as test set. Finally, the test set of BU-3DFE includes the two expressions Angry and Disgust, while the remeaining expressions were used for training.

We trained both Motion3DGAN and S2D-Dec using the Adam optimizer, with learning rate of 0.0001 and 0.001 and mini-batches of size 128 and 16, respectively. Motion3DGAN was trained for $8,000$ epochs, while 300 epochs were adopted for S2D-Dec. The hyper-parameters of the Motion3DGAN and S2D-Dec losses were set empirically to $\alpha_1 = 1$, $\alpha_2 = 10$, $\beta_1 = 1$ and $\beta_2 = 0.1$. We chose the mean SRVF of the CoMA data as a reference point p, where we defined the tangent space of $\mathcal{C}$.

4.2 3D Expression Generation: S2D-Dec

For evaluation, we set up a baseline by first comparing against standard 3DMM-based fitting methods. Similar to previous works [34], [53], we fit S^n to the set of target landmarks Z^e using the 3DMM components. Since the deformation is guided by the landmarks, we first need to select a corresponding set from S^n to be matched with Z^e. Given the fixed topology of the 3D faces, we can retrieve the landmark coordinates by indexing into the mesh, *i.e.*, $Z^n = S^n(\mathbf{I}_z)$, where $\mathbf{I}_z \in \mathbb{N}^n$ are the indices of the vertices that correspond to the landmarks. We then find the optimal deformation coefficients that minimize the Euclidean error between the target landmarks Z^e and the neutral ones Z^n, and use the coefficients to deform S^n. In the literature, several 3DMM variants have been proposed. We experimented the standard PCA-based 3DMM and the DL-3DMM in [53]. For fair comparison, we built the two 3DMMs using a number of deformation components comparable to the size of the S2D-Dec input, *i.e.*, $68 \times 3 = 204$. For PCA, we used either 38 components (99% of the variance) and 220, while for DL-3DMM we used 220 dictionary atoms.

With the goal of comparing against other deep models, we also considered the Neural3DMM [25]. It is a mesh auto-encoder tailored for learning a non-linear latent space of face variations and reconstructing the input 3D faces. In order to compare it with our model, we modified the architecture and trained the model to generate an expressive mesh S^g given its neutral counterpart as input. To do so, we concatenated the landmarks displacement (of size 204) to the latent vector (of size 16) and trained the network towards minimizing the same L_{pr} loss used in our model. We used the same data to train all the models for consistency. However, since we exclude identity reconstruction in our problem, it could be argued that multi-linear 3DMMs, where identity and

TABLE 1
Reconstruction error (mm) on expression-independent (left) and identity-independent (right) splits: comparison with PCA-k 3DMM (k components), DL-3DMM (220 dictionary atoms), and Neural3DMM.

	Expression Split				Identity Split			
Method	CoMA	D3DFACS	BU-3DFE	Florence 4D	CoMA	D3DFACS	BU-3DFE	Florence 4D
PCA-220	0.76 ± 0.73	0.42 ± 0.44	2.00 ± 1.67	0.70 ± 0.81	0.80 ± 0.73	0.56 ± 0.56	$\mathbf{2.10 \pm 1.74}$	0.16 ± 0.17
PCA-38	0.90 ± 0.84	0.44 ± 0.45	2.55 ± 1.72	0.87 ± 1.05	0.93 ± 0.82	0.58 ± 0.56	2.61 ± 1.78	0.18 ± 0.20
DL3DMM [53]	$0,86 \pm 0,80$	0.73 ± 1.15	2.09 ± 1.58	0.83 ± 1.03	0.89 ± 0.79	1.15 ± 1.50	2.22 ± 1.69	0.17 ± 0.18
Neural [25]	0.75 ± 0.85	0.59 ± 0.86	3.16 ± 2.12	1.45 ± 1.43	3.74 ± 2.34	2.09 ± 1.37	3.85 ± 2.32	1.41 ± 1.09
Ours	$\mathbf{0.52 \pm 0.59}$	$\mathbf{0.28 \pm 0.31}$	$\mathbf{1.97 \pm 1.52}$	$\mathbf{0.57 \pm 1.24}$	$\mathbf{0.55 \pm 0.62}$	$\mathbf{0.27 \pm 0.30}$	2.34 ± 1.83	$\mathbf{0.10 \pm 0.08}$

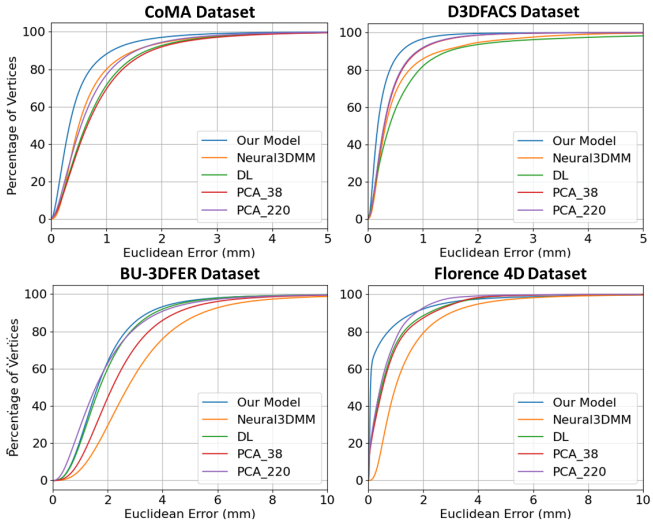


Fig. 5. Cumulative per-vertex error between PCA-based 3DMM models, DL-3DMM, Neural3DMM, and S2D-Dec, using expression-independent cross-validation on four datasets.

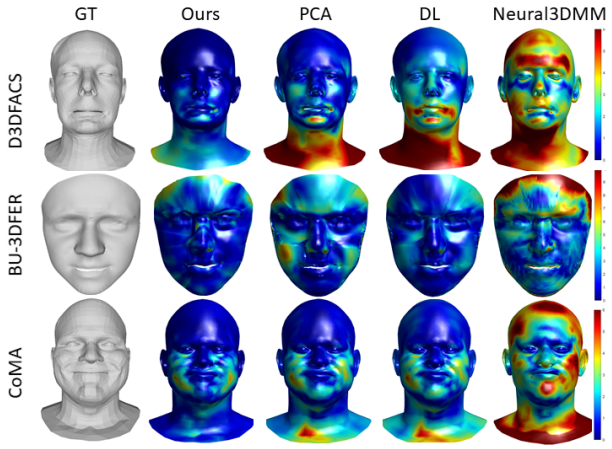


Fig. 6. Mesh reconstruction error (red=high, blue=low) of our model and other methods. Examples are given for three different datasets.

expressions are handled by two different models, should be used. We also experimented by building expression-specific 3DMMs, obtained by subtracting the neutral scan of each subject from their expressive counterparts instead of using the overall data mean. However, this not resulted in any noticeable improvement. Finally, we also identified the FLAME model [34]. Unfortunately, the training code of FLAME is not available, while using the model pre-trained on external data would result in an unfair comparison.

The mean per-vertex Euclidean error between the generated meshes and their ground truth is used as standard performance measure, as in the majority of works [3], [7], [20], [25]. Note that we exclude the Motion3DGAN model here as we do not have the corresponding ground-truth for the generated landmarks (they are generated from noise). Instead, we make use of the ground truth motion of the landmarks.

4.2.1 *Comparison with Other Approaches*

Table 1 shows a clear superiority of S2D-Dec over state-of-the-art methods for both the protocols and datasets, proving its ability to generate accurate expressive meshes close to the ground truth in both the case of unseen identities or expressions. In Figure 5, the cumulative per-vertex error distribution on the expression-independent splitting further highlights the precision of our approach, which can reconstruct 90%-98% of the vertices with an error lower than $1mm$. While other fitting-based methods retain satisfactory precision in both the protocols, we note that the performance of Neural3DMM [25] significantly drops when unseen identities are considered. This outcome is consistent to that reported in [20], in which the low generalization ability of these models is highlighted. Overall, our solution embraces the advantages of both approaches, being as general as fitting solutions yet more accurate. The only case where our method performs slightly worse is for the BU-3DFE dataset. Here, meshes are obtained through a dense registration, which is an error-prone process mostly for expressive scans. So, training data is likely affected by noise. However, results show S2D-Dec is quite robust.

Figure 13 shows some qualitative examples by reporting error heatmaps in comparison with PCA, DL-3DMM [53] and Neural3DMM [25] for the identity-independent splitting. The ability of our model as well as PCA and DL-3DMM to preserve the identity of the ground truth comes out clearly, in accordance with the results in Table 1. By contrast, Neural3DMM shows high error even for neutral faces, which proves its inability to generalize to the identity of unseen identities. Indeed, differently from the other methods, Neural3DMM encodes the neutral face in a latent space and predicts the 3D coordinates of the points directly. This evidences the efficacy of our S2D-Dec that learns per-point displacements instead of point coordinates.

4.2.2 *Transfer of Speech-Related Facial Movements*

By using landmarks, our S2D-Dec can transfer facial expressions or speech between identities. This is done by extracting the sequence of landmarks from the source face, encoding their motion

TABLE 2
Speech transfer reconstruction error on VOCASET

-	PCA	DL3DMM	Neural	Ours (S2D)
Rec. error (mm)	2.90	2.73	3.49	1.48

TABLE 3
Reconstruction error (mm): cross dataset setting

Method	Train: Florence 4D Test: CoMA	Train: CoMA Test: Florence 4D
PCA 220	1.81 ± 1.54	0.64 ± 0.89
DL	2.09 ± 1.78	0.72 ± 0.91
Ours	1.50 ± 1.84	0.56 ± 0.76

as an SRVF representation, transferring this motion to the neutral landmarks of the target face and using S2D-Dec to get the target identity following the motion of the first one. To demonstrate the high generalization ability of our proposed approach for different expressions, we evaluate it on VOCASET for speech transfer. This is done by transferring the speech-related movements from the first identity of VOCASET to the other 11 identities.

We report the reconstruction error between the obtained meshes after speech transfer and their ground truth counterparts. Note that the lengths of the ground truth and the obtained sequences are slightly different, thus we considered the error for each frame as the minimum error in a sliding windows of 20 frames centered on the given frame. In this experiment, we only considered the first five sentences of VOCASET that are shared between all identities. We highlight that the model used in this experiment was trained on the CoMA dataset that does not include such speech-related movements. That is possible given the full correspondence between CoMA and VOCASET data. Table 2 shows the results of this experiment. The superiority of our approach is clearly evidenced over other state-of-the-art solutions, which proves the high generalization ability of our method to animate 3D faces with completely different facial expressions from those seen during the training. In addition, these results demonstrate that our method can be used not only with our generated facial motions, but we can also exploit external motions that are completely different from our generated ones.

4.2.3 Cross Datasets Evaluation

We report the error obtained for a cross-dataset evaluation on two different datasets; COMA and Florence 4D. The error is reported on all CoMA samples and the test set of Florence 4D. In consistency with the previous results, Table 3 confirms the superiority of our model over other methods. We note that the mean errors obtained with the Florence 4D data are almost the same obtained with the expression split protocol in Table 1. However, a higher error is reached on CoMA with all methods.

4.2.4 Ablation Study

We report here an ablation study to highlight the contribution of each loss used to train S2D-Dec, with particular focus on our proposed weighted-$L1$ reconstruction loss. We conducted this study on the CoMA dataset using the first three identities as a testing set and training on the rest. This evaluation is based on the mean per-vertex error between the generated and the ground truth meshes. We evaluated three baselines, namely, $S1$, $S2$ and $S3$. For the first baseline ($S1$), we trained the model with the displacement reconstruction loss in (8) only. In $S2$, we added the standard $L1$ loss to $S1$, which corresponds to our loss in (9) without the landmark distance weights. To showcase the importance of weighting the contribution of each vertex, in $S3$ we added the landmark distance weights to the L_{pr} loss. Results are shown in Table 4, where the remarkable improvement of our proposed loss against the standard $L1$ turns out evidently. This is explained by the fact that assigning a greater weight to movable face parts allows the network to focus on regions that are subject to strong facial motions, ultimately resulting in realistic samples.

TABLE 4
Ablation study on the reconstruction loss of S2D-Dec

Method	Error (mm)
$S_1 : L_{dr}$	1.27 ± 1.88
$S_2 : S_1 + L_{pr}$ w/o distance weights	0.92 ± 1.33
$S_3 : S_1 + L_{pr}$	0.50 ± 0.56

4.3 4D Facial Expressions: Motion3DGan

We validated the performance of Motion3DGAN in a broad set of experiments, both quantitative and qualitative. However, since Motion3DGAN generates samples from noise to encourage diversity, the generated landmarks and meshes change at each forward pass. Thus, we cannot directly compute the mean per-vertex error with respect to ground-truth shapes as done in [3]. Comparing with other approaches is also not possible since no other method currently can generate dynamic transition sequences of arbitrary expressions. For a comprehensive analysis, we evaluated it in terms of *(i)* specificity error, and *(ii)* expression classification.

Specificity measure: Following the standard practice for statistical generative shape models, we use the *specificity* measure [54] to evaluate the quality of the generated samples. Given the very large number of possible start-end transitions (132 for the 12 expressions of CoMA), we selected a subset of them for validation. In particular, for each expression, we randomly chose 3 possible ending expressions, obtaining a total of 39 transitions. For each transition, we generated 64 samples (landmark sequences), for a total of $2,496$ samples, and computed the per-landmark average Euclidean distance with respect to the same transitions in the test data (as defined in Section 4.1). The average errors for all the cases are reported in Table 5. We first observe the error is stable and consistent across all the tested combinations. In addition, results show that transitions starting from the neutral expression score a lower distance. This is because the neutral expression is consistent across identities, while each subject performs facial expressions differently. In many cases, these can differ significantly; for example, some subjects of CoMA perform the "Eyebrows" expression by raising both of them, some others raise either the left or the right one (see Figure 3). Recalling (7), to obtain the landmarks from the generated motions, a reference landmark configuration for each expression needs to be chosen. Whereas for those starting from neutral this is not an issue, if the reference differs from that of the specific subject, the error might be larger even though the sequence is correct. To verify this, we performed a classification test, described in the next paragraph.

In Figure 7, the per-frame specificity error is reported. It can be observed as, even though the three sequences starting from

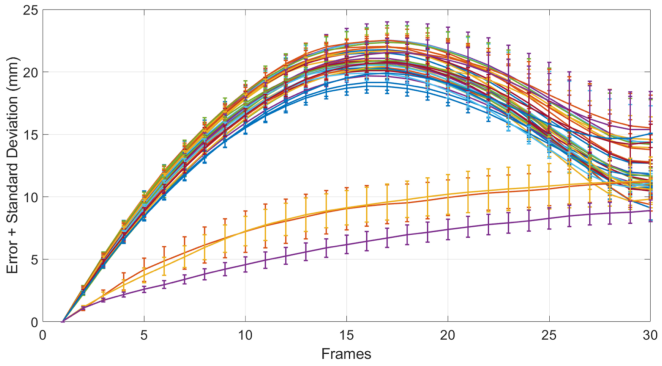


Fig. 7. Per-frame specificity error (mm) for all the 39 tested transitions. Legend is omitted for clarity. The three curves with lower error (purple, yellow, orange) refer to transitions starting from the neutral expression.

neutral obtain lower error on average, the error does not diverge. The higher increase in correspondence of the central frames (10-20) is again due to the very different and personal nature of facial expressions, which depends also on the velocity of performing it. The onset phase thus results more problematic, while at the peak of the expression (frame 30) the error tends to converge to a uniform value. However, the need of using an initial configuration of landmarks can be considered as a limitation of our approach; solving it would require generating also the starting configuration to ensure an even more pronounced diversity.

Expression classification: We further evaluated the quality of the generated sequences implementing a classification solution, similar to [3]. We trained a simple random forest classifier to recognize the 39 transitions generated in the previous paragraph. We trained this classifier on the same sequences used to train Motion3DGAN. For testing, we used the same $2,496$ samples. Since the S2D-Dec could compensate minor generation errors, we directly used the generated landmarks to perform classification.

Results are reported in Table 5, separately for each transition. Overall, the generated sequences have a high classification accuracy. meaning that they accurately resemble real ones, even though some of them score a lower accuracy. Qualitatively, we verified this is likely caused by the similarity of some classes of expressions in the CoMA dataset. For example, mouth extreme qualitatively looks similar to a combination of mouth open and mouth down, just differing in intensity.

Generating composed sequences: A novel characteristic of our method is that, even though the length of each transition is fixed to 30 frames, we are able to generate longer, composed and complex transitions. This is possible as we removed the constraint of starting the animation from a neutral face, and thanks to the SRVF representation, which allowed us to create interpolated transitions from one expression to another. So, it is possible for example to generate a 90 frames long sequence by composing three transitions, *e.g.*, neutral-bareteeth-eyebrows-lips up. To do so, we generate the sequence incrementally, using as starting landmark configuration for the $(i+1)$-th transition, the ending frame of the i-th one. To verify that the model is sufficiently robust to handle the diversity of each generated transition, we generated 64 samples of 5 composed sequences of 90 frames each (3 transitions). The average per-frame error is reported in Figure 8. Results show that the error does not significantly propagate across transitions, and remains quite stable even though a slight increasing trend is observed. This is due to the fact by switching from one expression peak to another without getting back to a neutral state, the resulting expression is actually a mix of the two. This is an interesting property, which makes the generated sequences even more natural looking. The lower peaks at frames 30 and 60 are instead due to the fact the training/testing sequences are 30 frames long, so leading to a discontinuity when computing the error. To clarify this aspect, let us consider the sequence "mouth down-mouth side-mouth open-lips back" (yellow curve in Figure 8): to compute the error from frame 0 to 30, we considered the corresponding transition in the real data; to compute the error for frames 30-60, we instead needed to consider a different transition, though the generated one starts from the last frame of the previous one. Ultimately, the discontinuity is reflected in the errors. Nonetheless, qualitative examples in Figure 9 and the supplementary video show the error propagation does not significantly corrupt the output.

TABLE 5
Specificity error (mm) and classification accuracy for 39 transitions

Start	End	Specificity (mm)	Classification
Bareteeth	High-Smile	15.1 ± 0.77	98%
	Lips-Back	15.3 ± 0.52	92%
	Lips-Up	14.4 ± 0.72	100%
Cheeks-In	Mouth-Down	16.4 ± 1.54	80%
	Mouth-Extreme	14.1 ± 0.41	86%
	Mouth-Middle	16.3 ± 0.13	75%
Eyebrow	Mouth-Open	14.7 ± 0.36	100%
	Mouth-Side	15.6 ± 0.72	91%
	Mouth-Up	16.1 ± 1.02	86%
High-Smile	Neutral	14.3 ± 0.36	100%
	Bareteeth	15.6 ± 0.72	100%
	Cheeks-In	16.1 ± 1.02	99%
Lips-Back	Eyebrow	14.3 ± 0.46	79%
	High-Smile	14.4 ± 0.59	98%
	Lips-Up	13.8 ± 0.59	100%
Lips-Up	Mouth-Down	16.0 ± 0.47	100%
	Mouth-Extreme	14.0 ± 1.09	81%
	Lips-Back	14.4 ± 0.45	76%
Mouth-Down	Mouth-Middle	16.4 ± 1.51	86%
	Mouth-Open	15.0 ± 0.40	95%
	Mouth-Side	15.6 ± 0.61	89%
Mouth-Extreme	Mouth-Up	15.9 ± 0.47	80%
	Neutral	14.0 ± 1.20	98%
	Bareteeth	14.5 ± 0.01	100%
Mouth-Middle	Cheeks-In	14.6 ± 0.56	100%
	Eyebrow	14.5 ± 0.47	75%
	High-Smile	14.6 ± 0.48	100%
Mouth-Open	Lips-Back	14.7 ± 0.45	100%
	Lips-Up	13.6 ± 0.54	100%
	Mouth-Down	16.0 ± 1.06	82%
Mouth-Side	Mouth-Extreme	15.2 ± 0.43	97%
	Mouth-Middle	15.9 ± 1.28	78%
	Mouth-Open	15.0 ± 0.43	100%
Mouth-Up	Mouth-Side	15.1 ± 0.52	88%
	Neutral	13.4 ± 0.75	100%
	Bareteeth	14.1 ± 0.08	100%
Neutral	Bareteeth	7.9 ± 1.32	100%
	Cheeks-In	8.0 ± 1.28	94%
	Eyebrow	5.7 ± 0.62	100%

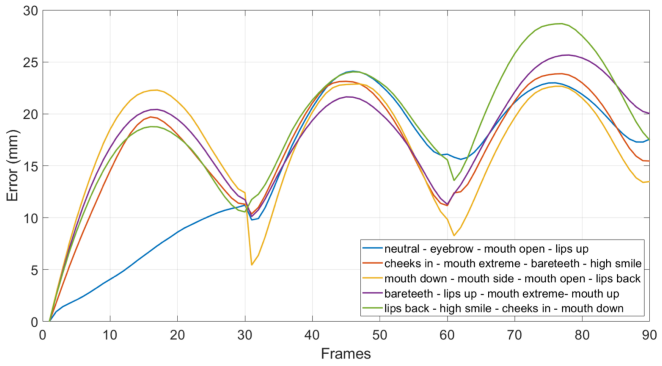


Fig. 8. Per-frame evolution of the error for 5 sequences composed of 3 transitions each. Each transition is 30 frames long.

4.4 Subjective Evaluation

In this section, we report the results of a user study aimed at assessing the perceived quality of the generated 4D expression sequences. We recruited 23 participants, and asked them to evaluate the quality of the generated data in terms of "Naturalness" (NAT) score, ranging from 0 to 10. To obtain consistent results, before starting the questionnaire, users were shown two reference examples for the lower and upper bounds of the score; 0 corresponds to a completely failed animation, while 10 corresponds to a real sequence from the original dataset. In addition, we also asked the users to evaluate the faithfulness of a given animation with respect to the input label (LAB). The latter score also ranges from 0 to 10. Users were shown a total of 28 generated animations, 14 of which included multiple transitions, and 14 only neutral-peak transitions.

The average reported NAT score was of 7.22, while the LAB score of 7.05, indicating users perceived the generated meshes as fairly natural and faithful to the input label. Concerning the neutral-peak sequences, they were generated both using the proposed model, and that of [4]. The NAT score reported for our model was of 7.39, and 5.82 for the model of [4]. Our results were perceived as more realistic, possibly thanks to the increased variability and quantity of the training data, whose collection was made possible by augmenting them with interpolated samples.

4.5 Qualitative Results

Figure 9 shows additional qualitative results. In particular, Figure 9 (top) shows two examples of composed sequences generated with our model. These are obtained by generating transitions between different expressions incrementally with Motion3DGAN. S2D-Dec was applied to these motions ultimately obtaining a complex 4D sequence. Figure 9 (bottom), instead, shows some qualitative examples of speech transfer, applied as described in Section 4.2.2. The reader can appreciate that our model allows a natural transfer of speech related movements that were completely unseen during training.

5 Conclusions and Limitations

In this paper, we proposed a novel framework for dynamic 3D facial expression generation. From a starting 3D face and an expression label indicating the starting and the ending expression, we can synthesize sequences of 3D faces switching between different facial expression. This is achieved by two decoupled networks that separately address the motion dynamics modeling and generation of expressive 3D faces from a starting one. We demonstrated the improvement with respect to previous literature, the high generalization ability of the model to unseen expressions and identities, and showed that using landmarks is effective in modeling the motion of expressions and the generation of 3D meshes. We also identified two main limitations: first, our S2D-Dec generates expression-specific deformations, and so cannot model identities. Moreover, while Motion3DGAN can generate diverse expressions including transition between expressions and long composed 4D expressions, the samples are of a fixed length.

6 Acknowledgments

This work was supported by the French State, managed by National Agency for Research (ANR) National Agency for Research (ANR) under the Investments for the future program with reference ANR-16-IDEX-0004 ULNE. This paper was also partially supported by the European Union's Horizon 2020 research and innovation program under grant number 951911 - AI4Media. The authors would also like to thank Giulio Calamai for performing part of the experimentation. Most of this work was done when Naima Otberdout was at the University of Lille.

7 Appendix

7.1 Landmarks Configuration

In Figure 10 we show, for three different expressions, the configuration of landmarks used to guide the generation of the facial expression.

7.2 Logarithm and Exponential Maps

In order to map the SRVF data forth and back to a tangent space of $\mathcal{C}$, we use the logarithm $\log_p(.)$ and the exponential $\exp_p(.)$ maps defined in a given point p by,

$$\begin{aligned} \log_p(q) &= \frac{d_{\mathcal{C}}(q,p)}{sin(d_{\mathcal{C}}(q,p))}(q - cos(d_{\mathcal{C}}(q,p))p), \\ \exp_p(s) &= cos(\|s\|)p + sin(\|s\|)\frac{s}{\|s\|}, \end{aligned} \tag{11}$$

where $d_{\mathcal{C}}(q,p) = \cos^{-1}(\langle q,p\rangle)$ is the distance between q and p in $\mathcal{C}$.

7.3 Architecture of S2D-Dec

The architecture adopted for S2D-Dec is based on the architecture proposed in [25]. S2D-Dec takes as input the displacements of 68 landmarks illustrated in Figure 10. The architecture includes a fully connected layer of size 2688, five spiral convolution layers of 64, 32, 32, 16 and 3 filters. Each spiral convolution layer is followed by an up-sampling by a factor of 4.

7.4 Ablation Study

In this section, we report a visual comparison between reconstructions obtained with the standard L1 loss and our proposed weighted L1. Figure 12 clearly shows the effect of our introduced weighting scheme that allows for improved expression modeling.

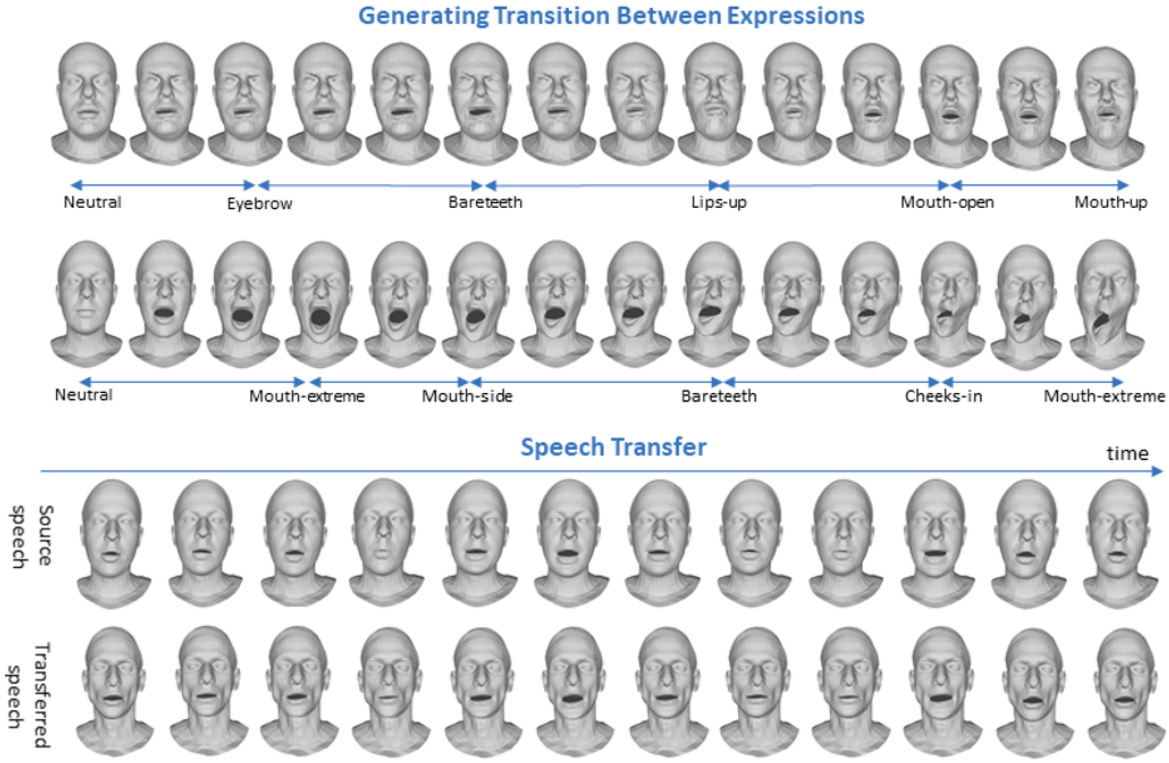


Fig. 9. **Applications** – From top to bottom: **Transition between expressions**: starting from a neutral face, each row show the transition between five different facial expressions. **Transfer**: speech transfer from one identity to another. Animated versions can be found in this link.

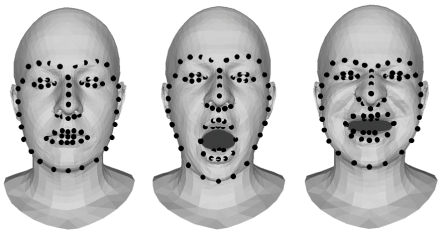


Fig. 10. Landmarks configuration used to guide our model.

REFERENCES

[1] L. Fan, W. Huang, C. Gan, J. Huang, and B. Gong, "Controllable image-to-video translation: A case study on facial expression generation," in *Conf. on Artificial Intelligence (AAAI) Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2019, pp. 3510–3517.

[2] N. Otberdout, M. Daoudi, A. Kacem, L. Ballihi, and S. Berretti, "Dynamic facial expression generation on hilbert hypersphere with conditional wasserstein generative adversarial nets," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 2, pp. 848–863, 2022.

[3] R. A. Potamias, J. Zheng, S. Ploumpis, G. Bouritsas, E. Ververas, and S. Zafeiriou, "Learning to generate customized dynamic 3D facial expressions," in *European Conf. on Computer Vision (ECCV)*, 2020, pp. 278–294.

[4] N. Otberdout, C. Ferrari, M. Daoudi, S. Berretti, and A. Del Bimbo, "Sparse to dense dynamic 3D facial expression generation," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 20 385–20 394.

[5] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, and P. Liu, "A high-resolution spontaneous 3d dynamic facial expression database," in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 2013, pp. 1–6.

[6] D. Cosker, E. Krumhuber, and A. Hilton, "A facs valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling," in *IEEE Int. Conf. on Computer Vision*. IEEE, 2011, pp. 2296–2303.

[7] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, "Generating 3D faces using convolutional mesh autoencoders," in *European Conf. on Computer Vision (ECCV)*, 2018, pp. 725–741.

[8] C. Cao, Q. Hou, and K. Zhou, "Displaced dynamic expression regression for real-time facial tracking and animation," *ACM Trans. on Graphics*, vol. 33, no. 4, Jul. 2014.

[9] D. Cudeiro, T. Bolkart, C. Laidlaw, A. Ranjan, and M. J. Black, "Capture, learning, and synthesis of 3D speaking styles," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 093–10 103.

[10] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Trans. on Graphics*, vol. 36, no. 4, Jul. 2017.

[11] X. Ji, H. Zhou, K. Wang, Q. Wu, W. Wu, F. Xu, and X. Cao, "Eamm: One-shot emotional talking face via audio-based emotion-aware motion model," in *ACM SIGGRAPH 2022 Conference Proceedings*, 2022, pp. 1–10.

[12] Z. Zhao, Y. Zhang, T. Wu, H. Guo, and Y. Li, "Emotionally controllable talking face generation from an arbitrary emotional portrait," *Applied Sciences*, vol. 12, no. 24, p. 12852, 2022.

[13] S. E. Eskimez, Y. Zhang, and Z. Duan, "Speech driven talking face generation from a single image and an emotion condition," *IEEE Transactions on Multimedia*, vol. 24, pp. 3480–3490, 2021.

[14] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Annual Conf. on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 1999, pp. 187–194.

[15] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway, "A 3D morphable model learnt from 10,000 faces," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5543–5552.

[16] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, 2009, pp. 296–301.

[17] A. Brunton, T. Bolkart, and S. Wuhrer, "Multilinear wavelets: A statistical shape space for human faces," in *European Conf. on Computer Vision*. Springer, 2014, pp. 297–312.

[18] T. Neumann, K. Varanasi, S. Wenger, M. Wacker, M. Magnor, and

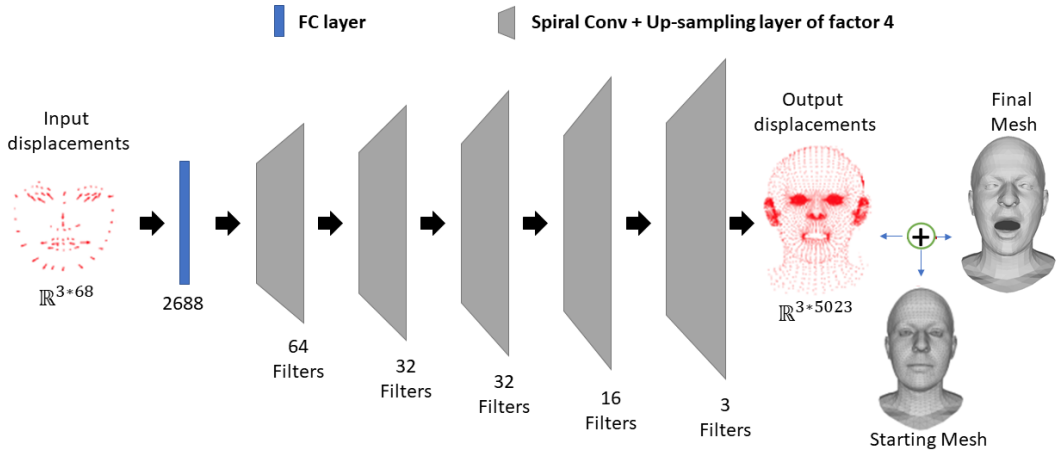


Fig. 11. Architecture of the Sparse2Dense decoder (S2D-Dec).

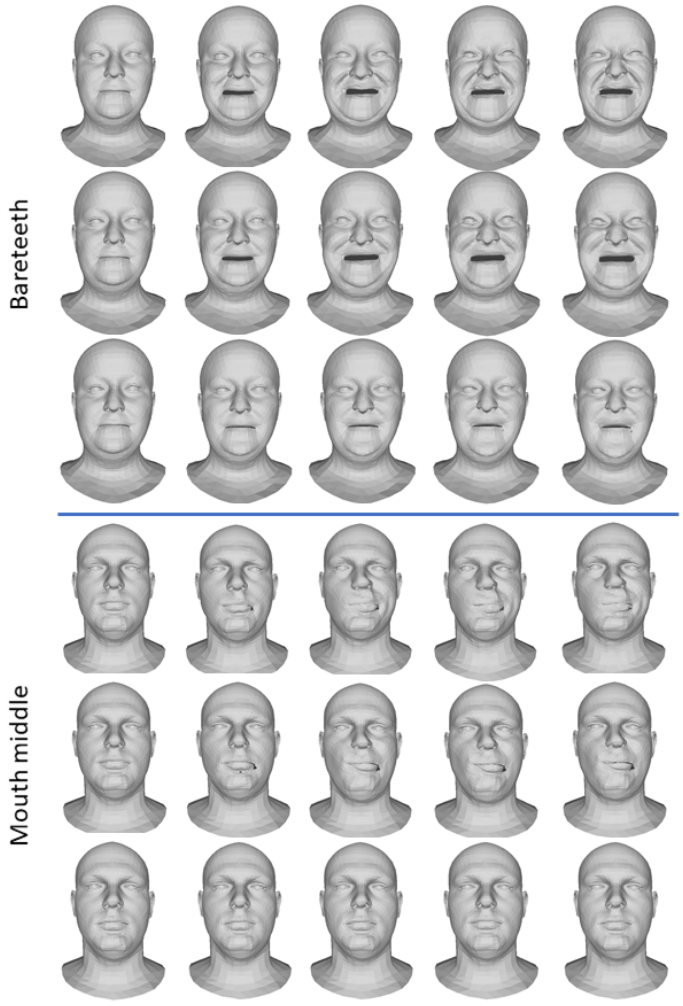


Fig. 12. Ablation study: qualitative comparison between ground truth (first row) our model with (second row) and without (last row) weighted loss.

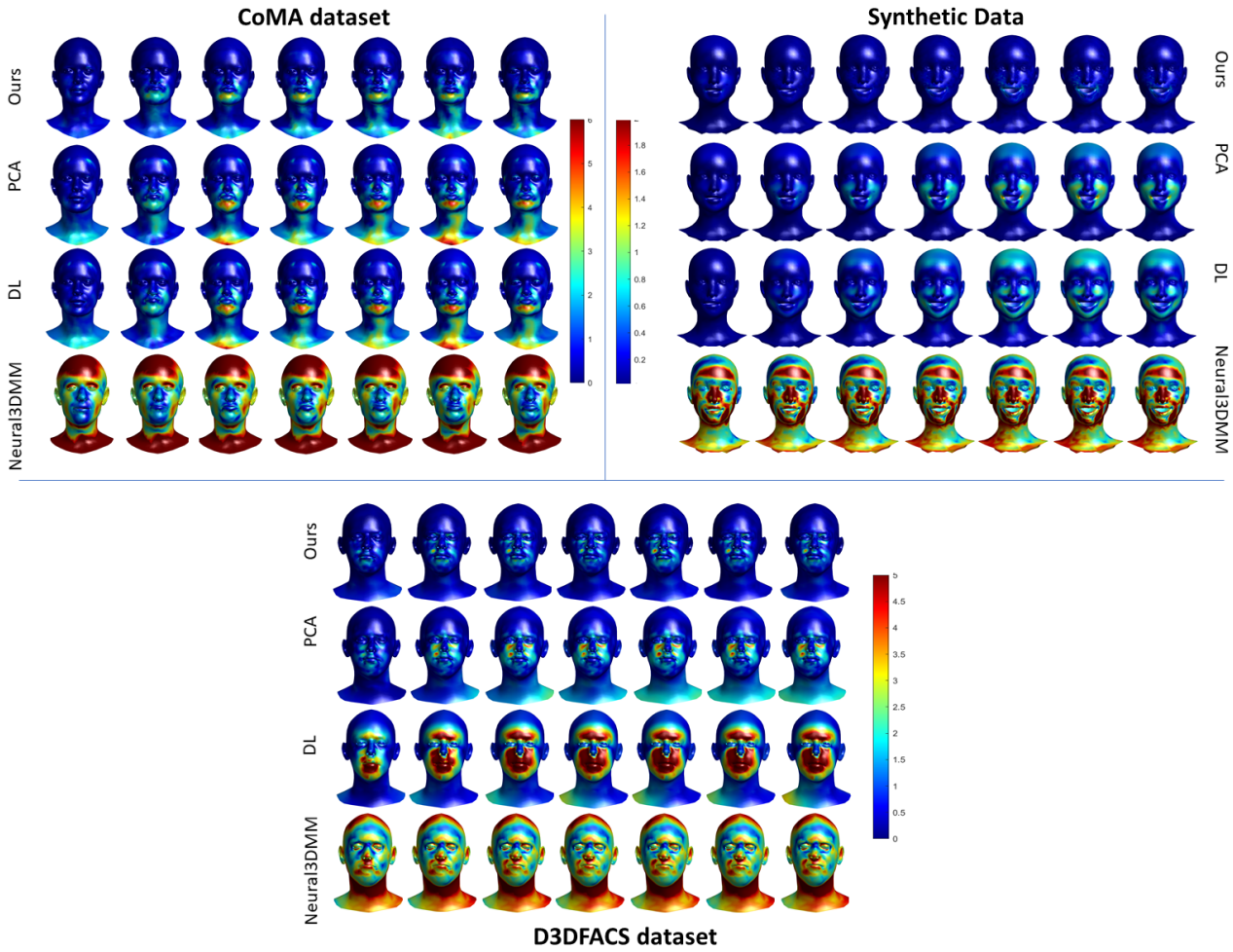


Fig. 13. Temporal evolution of the mesh reconstruction error (red=high, blue=low) from the neutral face to the apex expression of our model and other methods. Examples from three different databases.

C. Theobalt, "Sparse localized deformation components," *ACM Trans. on Graphics (TOG)*, vol. 32, no. 6, pp. 1–10, 2013.

[19] M. Lüthi, T. Gerig, C. Jud, and T. Vetter, "Gaussian process morphable models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 40, no. 8, pp. 1860–1873, 2017.

[20] C. Ferrari, S. Berretti, P. Pala, and A. Del Bimbo, "A sparse and locally coherent morphable face model for dense semantic correspondence across heterogeneous 3D faces," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2021.

[21] C. Ferrari, G. Lisanti, S. Berretti, and A. Del Bimbo, "Dictionary learning based 3D morphable model construction for face recognition with varying expression and pose," in *Int. Conf. on 3D Vision*. IEEE, 2015, pp. 509–517.

[22] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: Going beyond euclidean data," *IEEE Signal Processing Mag.*, vol. 34, no. 4, pp. 18–42, 2017.

[23] O. Litany, A. Bronstein, M. Bronstein, and A. Makadia, "Deformable shape completion with graph convolutional autoencoders," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2018, pp. 1886–1895.

[24] N. Verma, E. Boyer, and J. Verbeek, "Feastnet: Feature-steered graph convolutions for 3D shape analysis," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2018, pp. 2598–2606.

[25] G. Bouritsas, S. Bokhnyak, S. Ploumpis, S. Zafeiriou, and M. Bronstein, "Neural 3D morphable models: Spiral convolutional networks for 3D shape representation learning and generation," in *IEEE/CVF Int. Conf. on Computer Vision (ICCV)*, 2019, pp. 7212–7221.

[26] I. Lim, A. Dielen, M. Campen, and L. Kobbelt, "A simple approach to intrinsic correspondence learning on unstructured 3D meshes," in *European Conf. on Computer Vision (ECCV) Workshops*, 2018.

[27] V. F. Abrevaya, S. Wuhrer, and E. Boyer, "Multilinear autoencoder for 3D face model learning," in *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2018, pp. 1–9.

[28] S. Moschoglou, S. Ploumpis, M. A. Nicolaou, A. Papaioannou, and S. Zafeiriou, "3dfacegan: Adversarial nets for 3D face representation, generation, and translation," *Int. Journal of Computer Vision*, vol. 128, pp. 2534–2551, 2020.

[29] L. Chen, H. Su, and Q. Ji, "Deep structured prediction for facial landmark detection," in *Advances in Neural Information Processing Systems (Neurips)*, vol. 32, 2019.

[30] X. Dong, Y. Yang, S.-E. Wei, X. Weng, Y. Sheikh, and S.-I. Yu, "Supervision by registration and triangulation for landmark detection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[31] J. Wan, Z. Lai, J. Li, J. Zhou, and C. Gao, "Robust facial landmark detection by multiorder multiconstraint deep networks," *IEEE Trans. on Neural Networks and Learning Systems*, pp. 1–14, 2021.

[32] S. Z. Gilani, A. Mian, and P. Eastwood, "Deep, dense and accurate 3d face correspondence for generating population specific deformable models," *Pattern Recognition*, vol. 69, pp. 238–250, 2017.

[33] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: A 3D total solution," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 78–92, 2017.

[34] T. Li, T. Bolkart, M. Julian, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans," *ACM Trans. on Graphics, (Proc. SIGGRAPH Asia)*, vol. 36, no. 6, 2017.

[35] C. Ferrari, S. Berretti, P. Pala, and A. Del Bimbo, "Rendering realistic subject-dependent expression images by learning 3dmm deformation coefficients," in *European Conf. on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[36] P. Garrido, L. Valgaerts, O. Rehmsen, T. Thormahlen, P. Perez, and C. Theobalt, "Automatic face reenactment," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 2014, pp. 4217–4224.

[37] F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni, "Expnet: Landmark-free, deep, 3d facial expressions," in *IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG)*. IEEE, 2018, pp. 122–129.

[38] F.-J. Chang, A. Tuan Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni, "Faceposenet: Making a case for landmark-free face alignment," in *IEEE Int. Conf. on Computer Vision Workshops*, 2017, pp. 1599–1608.

[39] B. Gecer, S. Ploumpis, I. Kotsia, and S. Zafeiriou, "Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2019, pp. 1155–1164.

[40] W. Wang, X. Alameda-Pineda, D. Xu, P. Fua, E. Ricci, and N. Sebe, "Every smile is unique: Landmark-guided diverse smile generation," in *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2018, pp. 7083–7092.

[41] D. Zeng, H. Liu, H. Lin, and S. Ge, "Talking face generation with expression-tailored generative adversarial network," in *ACM Int. Conf. on Multimedia (MM'20)*, 2020, p. 1716–1724.

[42] A. Kacem, M. Daoudi, B. Ben Amor, and J. Carlos Alvarez-Paiva, "A novel space-time representation on the positive semidefinite cone for facial expression recognition," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2017, pp. 3180–3189.

[43] A. Srivastava, E. Klassen, S. H. Joshi, and I. H. Jermyn, "Shape analysis of elastic curves in euclidean spaces," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 7, pp. 1415–1428, 2011.

[44] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo, "3-D human action recognition by shape analysis of motion trajectories on Riemannian manifold," *IEEE Trans. on Cybernetics*, vol. 45, no. 7, pp. 1340–1352, 2014.

[45] H. Drira, B. Ben Amor, A. Srivastava, M. Daoudi, and R. Slama, "3D face recognition under expressions, occlusions, and pose variations," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 35, no. 9, pp. 2270–2283, 2013.

[46] Z. Huang, J. Wu, and L. Van Gool, "Manifold-valued image generation with wasserstein generative adversarial nets," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI*. AAAI Press, 2019, pp. 3886–3893.

[47] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 5767–5777.

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

[49] F. Principi, S. Berretti, C. Ferrari, N. Otberdout, M. Daoudi, and A. Del Bimbo, "The florence 4d facial expression dataset," in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2023, pp. 1–6.

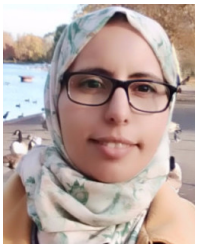
[50] I. Daz Productions. (2022) Daz 3D. [Online]. Available: https://www.daz3d.com/

[51] A. D. Bagdanov, A. Del Bimbo, and I. Masi, "The florence 2d/3d hybrid face dataset," in *Joint ACM Workshop on Human Gesture and Behavior Understanding*, 2011, p. 79–80.

[52] E. Pierson, M. Daoudi, and A.-B. Tumpach, "A riemannian framework for analysis of human body surface," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2991–3000.

[53] C. Ferrari, G. Lisanti, S. Berretti, and A. Del Bimbo, "A dictionary learning-based 3D morphable shape model," *IEEE Trans. on Multimedia*, vol. 19, no. 12, pp. 2666–2679, 2017.

[54] R. Davies, C. Twining, and C. Taylor, *Statistical models of shape: Optimisation and evaluation*. Springer Science & Business Media, 2008.



Naima Otberdout is currently a Postdoctoral researcher in the University of Lille, France. She received the master's degree in computer sciences and telecommunication from Mohammed V University, Rabat, Morocco in 2016. She received the Ph.D. degree in computer science from the same university in 2021. Her current research interests include computer vision and pattern recognition with applications to human behavior understanding.

Claudio Ferrari is currently assistant professor at the department of Architecture and Engineering of the University of Parma. He received the Ph.D. in Information Engineering from the University of Florence, in 2018. He has been a visiting research scholar at the University of Southern California in 2014. His research interest focus on 3D/4D face and body analysis, human emotion, biometrics and behavior understanding.

Mohamed Daoudi is Full Professor of Computer Science at IMT Nord Europe and the lead of Image group at CRIStAL Laboratory (UMR CNRS 9189). His research interests include pattern recognition, shape analysis and computer vision. He has published over 150 papers in some of the most distinguished scientific journals and international conferences. He is/was Associate Editor of Image and Vision Computing Journal, IEEE Trans. on Multimedia, Computer Vision and Image Understanding, IEEE Trans. on Affective Computing and Journal of Imaging. He has served as General Chair of IEEE International Conference on Automatic Face and Gesture Recognition, 2019. Prof. Daoudi is IAPR Fellow.

Stefano Berretti is an Associate Professor at University of Florence, Italy. He has been Visiting Professor at University of Lille, and University of Alberta. His research interests focus on 3D computer vision for face biometrics, human emotion and behavior understanding. On these themes he published more than 200 journals and conference papers. He is an Associate Editor of ACM TOMM, IEEE TCSVT, and of the IET *Computer Vision journal*.

Alberto Del Bimbo is Full Professor of Computer Engineering at the University of Florence, Italy. His scientific interests include multimedia retrieval, pattern recognition, image and video analysis and human–computer interaction. Prof. Del Bimbo is IAPR Fellow, Associate Editor of several journals in the area of pattern recognition and multimedia, and the Editor-in-Chief of the ACM Transactions on Multimedia Computing, Communications, and Applications. He was also the recipient of the prestigious SIGMM 2016 Award for Outstanding Technical Contributions to Multimedia Computing, Communications and Applications.