EUROGRAPHICS 2020 / U. Assarsson and D. Panozzo
(Guest Editors)

Volume 39 (2020), Number 2

Facial Expression Synthesis using a Global-Local Multilinear Framework

M. Wang[1,2] D. Bradley[1] S. Zafeiriou[1,2] T. Beeler[1]

[1]DisneyResearch|Studios [2]Imperial College London

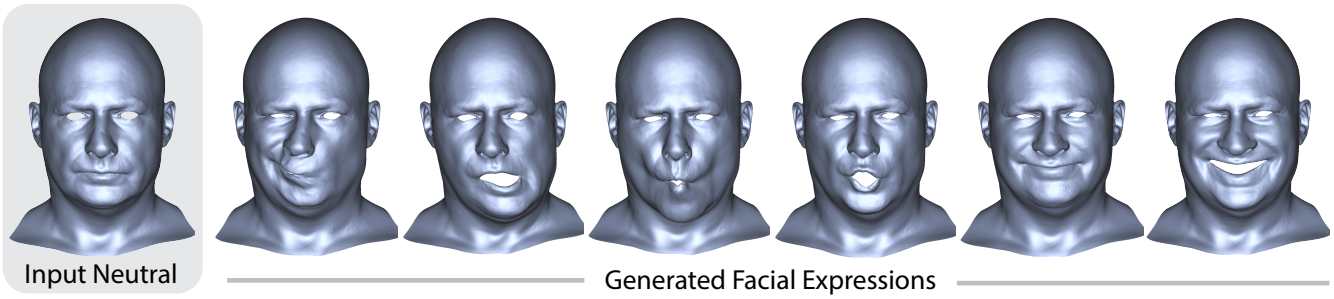

Figure 1: *Given a neutral face mesh as input, our method can synthesize highly-realistic 3D facial expressions that are plausible for the target subject and are free of the common artifacts of current state-of-the-art methods.*

Abstract

We present a practical method to synthesize plausible 3D facial expressions for a particular target subject. The ability to synthesize an entire facial rig from a single neutral expression has a large range of applications both in computer graphics and computer vision, ranging from the efficient and cost-effective creation of CG characters to scalable data generation for machine learning purposes. Unlike previous methods based on multilinear models, the proposed approach is capable to extrapolate well outside the sample pool, which allows it to plausibly predict the identity of the target subject and create artifact free expression shapes while requiring only a small input dataset. We introduce global-local multilinear models that leverage the strengths of expression-specific and identity-specific local models combined with coarse motion estimations from a global model. Experimental results show that we achieve high-quality, plausible facial expression synthesis results for an individual that outperform existing methods both quantitatively and qualitatively.

CCS Concepts

• ***Computing methodologies*** → ***Computer vision representations; Shape modeling;***

1. Introduction

Most approaches for facial animation [DN08], model-based face reconstruction [ZTG*18], or facial performance capture employ some form of blendshapes [LAR*14] as a deformation subspace. For high-quality facial rigs it is not uncommon to use hundreds or even thousands of blendshapes for a single person, e.g. `https://www.3lateral.com`. Building such a rig is highly involved, and typically requires to capture a human subject under a large range of expressions, or to manually sculpt these shapes when creating an imaginary person. Hence these high-quality person-specific facial rigs are currently only practical for hero assets in high budget feature films.

On the other end of the spectrum, consumer-based facial capture methods typically employ pre-built multilinear blendshape models rather than person-specific ones, oftentimes referred to as morphable models [BV*99]. In order to remain generic, however, these multilinear models must contain many identities with a shape per expression per identity, and hence require even more shapes to be acquired. As a result, these multi-identity models typically contain only a very limited number of expressions or identities, and as a consequence inherently limit the quality that may be achieved.

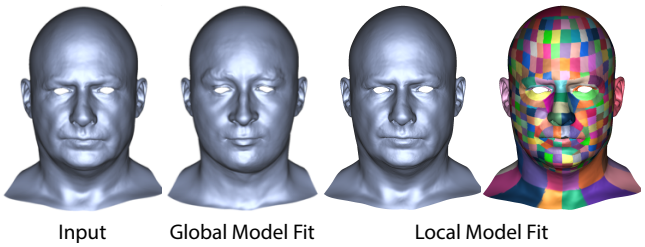

Figure 2: *Model fitting to an input neutral face (left), using a traditional global face model (center-left) and using our new global-local framework (right). Unlike the global model, our proposed approach is able to extrapolate **outside** of the training set and better recover the identity of the input subject, even though trained on the same dataset.*

When capturing or manually sculpting shapes is not an option, and improved quality over morphable models is desired, then shape synthesis can provide a viable alternative. Starting from a single neutral scan or sculpt of the target subject, a synthesis technique can generate the corresponding expression meshes automatically. A good synthesis algorithm should satisfy several criteria - 1) the identity of the target subject must be maintained within all synthesized expressions, 2) the generated expressions should contain plausible deformation at both the global and local scale, and 3) the resulting meshes should be free from geometric artifacts like local pinching, abnormal stretching and triangle flipping. Previous methods for synthesizing facial expressions tend to fall short on one or more of these criteria. Our proposed method can synthesize the full set of expression blendshapes while generally satisfying all criteria, which we will demonstrate both quantitatively as well as qualitatively through a user study.

Our method is data-driven and based on a limited number of samples (∼50 people), but unlike traditional multilinear models it is able to create plausible shapes that lie **outside** of this dataset since it combines a global tensor model to synthesize the coarse deformation with local patch-wise models to generate the detailed shape deformation, visualized in Figure 2. The novelty of our method lies in a) the combination of global and local multilinear methods and b) the separation of the local identity and local expression models. The local expression models ensure plausible expressions and the local identity models aim to preserve identity across expressions. The final synthesized shapes exhibit expression specific detail while being plausible for the identity of the subject. This allows on the one hand to effortlessly create high-quality facial rigs also for background characters or lower budget productions, and on the other hand to augment existing multi-identity models such as [BRZ*16] with expressions, which will in turn benefit methods that leverage multilinear morphable models for fitting and performance capture.

2. Related Work

2.1. Expression Synthesis

Facial expression synthesis has been an active research topic. Prior work can be mainly summarized in two categories. The first category focuses on generative models such as multilinear models whereas the second category comprises of computer graphics techniques which directly warp input faces to target expressions. We split the discussion into 2D and 3D expression synthesis.

2.1.1. 2D Expression Synthesis

In 2D, a lot of work has been completed on facial expression synthesis. Especially, Generative Adversarial Networks (GANs) have shown a lot of promise for this task. StarGAN [CCK*18] displays successful results for facial expression synthesis by conditioning the generation with images of a specific domain. Here, such a domain would be a set of images of persons sharing the same expression. Pumarola et al [PAM*18] proposed an unsupervised extension of this using Action Units (AUs). Other works [SYH*17, WSC*19] have investigated the potential of incorporating 3D models in editing the expression of the face in 2D. Our work targets 3D expression synthesis.

2.1.2. 3D Expression Synthesis

By carefully designing a data tensor according to modes of variation such as identity and expression, TensorFaces [VT02] is able to analyze each mode linearly, since each mode is allowed to vary in turn, while the remaining modes are held constant. A new identity of a given expression can be projected into the core tensor by fixing the expression mode. Once the identity component is estimated, the identity mode can be fixed while the expression mode can be varied to generate new expressions for the input identity.

Local methods based on multilinear models have been proposed by [BBW14]. Though they outperform traditional mulitlinear methods in capturing details, their ability to synthesize high-quality, artifact free expressions is still limited.

Blanz and Vetter [BV*99] proposed 3D Morphable Models (3DMMs) to model the shape and texture of the human face. Based on this, Blanz et al. [BBPV03] transferred expressions from a common expression framework to reanimate a face in a still image or video. Later, Vlasic et al. [VBPP05] proposed multilinear face models for facial expression tracking and transfer. In order to construct their multilinear models from an incomplete set of face scans they applied an expectation-maximization approach to fill in the missing data. Wang et al. [WPSZ18] proposed an unsupervised method to build a multilinear model from partial data using a custom tensor decomposition. FaceWarehouse [CWZ*14] introduced a 3D facial expression database. They built a bilinear face model from the data and showed that it can be used to estimate face identities and expressions for facial images and videos. Recently [LBB*17] introduced a linear model of expression, which they trained on very limited number of people (∼11) on a low-dimensional head model. A common problem with all current morphable model approaches is that they require a tremendous amount of data in order to be flexible and generic, due to their global nature.

Expression cloning, where one person's expression is transferred onto another person's neutral face, is a popular technique for synthesizing facial expressions in computer graphics. Sumner and Popovic [SP04] proposed nonlinear deformation transfer to transfer

the 3D deformations from a source mesh to a target mesh. An elastic model was proposed by [ZLZ*14] to balance the global and local warping effects aimed at 2D facial expression synthesis. While approaches based on traditional multilinear models leverage the whole dataset and estimate plausible expressions for a person but do not capture details well, expression cloning approaches do not attempt to reflect the expression of the target person but can achieve high-quality expression synthesis by simply cloning the expression details of the source. This avoids the need for a large dataset but constrains the expression details to be a copy of the source person. In this work, we propose a novel method that leverages the advantages of multilinear models in predicting plausible expressions for a target subject and at the same time achieves high-quality synthesis that produces plausible expression details for an individual.

2.2. 3D Face Datasets

In recent years, various databases have been collected for expressive 3D faces. One of the earliest, BU-3DFE [YWS*06] contains 100 subjects over 7 expressions (neutral and the six prototypic expressions). More recently 4D facial expression databases such as BU-4DFE [YCS*08] and 4DFAB [CKPZ18] have been released. 4DFAB [CKPZ18] is currently the largest with 180 people captured. Despite these big efforts, the amount of available 3D facial expression data is still limited, hindering the performance of multilinear models built from them. On the other hand, 3D face datasets containing only neutral scans are often several orders of magnitude larger. Booth et al. [BRZ*16], for example, collected 10,000 faces to build a 3D morphable face model. Our proposed method would be able to augment these datasets with expressions.

3. Method Overview

Given a 3D face mesh of a target subject with neutral expression, our goal is to produce realistic deformations of the input mesh that correspond to different facial expressions of the target subject. We take a data-driven approach, and construct a novel global-local multilinear model for expression synthesis, built from a small corpus of capture data of 3D facial expressions. A schematic view of our method is given in Figure 3.

Traditional multilinear models [VT02, VBPP05] typically consider the face globally, assuming that any new sample (i.e. face mesh) may be encoded as a linear combination of the basis vectors of the tensor. While this assumptions holds reasonably well for lower resolution geometry, it does not scale. For higher resolution shapes, the dimensionality of the problem becomes too large and new samples lie far outside of the convex hull spanned by the data tensor. Still, we argue that even if the sample can only be approximated at a coarse level, the coarse deformation will be reasonable, and thus we also begin with a traditional global multilinear model fit. In order to also synthesize believable fine scale expression deformation and to faithfully represent the identity of the subject, we suggest to leverage a set of local multilinear models. Like a global model, each of them encodes shape and deformation variation, but since they are highly localized they need to encode much less variability. While solving within the local models, identity and expression could be optimized jointly, however we argue that it is better to solve them independently as this gives the model more expressive power since we are not enforcing explicit correlation between identity and expression dimensions. Hence we extract two linear slices from the local multilinear models, where the first slice yields *Local Identity Models* that contain expression variation for a specific identity, and the second slice results in *Local Expression Models* that contain identity variation for a specific expression.

Our approach to combine the global and local models is to first fit the global model to the input neutral mesh, and subsequently pose the model into a desired expression k. Sparsely sampling the resulting shape at sample locations inspired by mocap marker layouts yields the coarse deformation constraints used to fit the local expression models. This results in a plausible expression shape free from artifacts, yet is not guaranteed to approximate the identity of the target subject. Hence we densely sample the expression shape and fit the local identity models to the resulting constraints, aiming to reproduce the expression shape as closely as possible within the identity subspace, yielding a final expression that preserves both the predicted identity and expression details.

It is important to note that this is not a hierarchical setup, where the local tensors encode purely the residual. The local models can fully describe the face without requiring the global model and they are not computed based on the residual of the global model. The global model provides only cues for coarse deformation. The advantage of our global-local approach over a pure hierarchical setup is that the local models can correct eventual errors made by the global model.

We begin by introducing important mathematical notations (Section 4) and outline the dataset that our method employs (Section 5) before presenting the approach in Section 6. Section 7 provides a detailed evaluation of our method compared to existing state-of-the-art approaches.

4. Notations

Throughout the paper, matrices are denoted by uppercase boldface letters (e.g. $\boldsymbol{X}$), vectors by lowercase boldface letters (e.g. $\boldsymbol{x}$). Tensors are considered as the multidimensional equivalent of matrices (second-order tensors) and vectors (first-order tensors) and are denoted by calligraphic letters (e.g. $\mathcal{X}$). The *order* of a tensor is the number of indices needed to address its elements. Consequently, each element of an Mth-order tensor $\mathcal{X}$ is addressed by M indices, i.e. $(\mathcal{X})_{i_1 i_2 \ldots i_M} = x_{i_1 i_2 \ldots i_M}$. The sets of real and integer numbers are denoted by $\mathbb{R}$ and $\mathbb{Z}$, respectively. An Mth-order real-valued tensor $\mathcal{X}$ is defined over the tensor space $\mathbb{R}^{I_1 \times I_2 \times \cdots \times I_M}$, where $I_m \in \mathbb{Z}$ for $m = 1, 2, \ldots, M$. The mode-m (matrix) product of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_M}$ with a matrix $\boldsymbol{U} \in \mathbb{R}^{J \times I_m}$ is denoted by $\mathcal{X} \times_m \boldsymbol{U} \in \mathbb{R}^{I_1 \times \cdots \times I_{m-1} \times J \times I_{m+1} \times \cdots \times I_M}$. Element-wise, it is defined as

$$(\mathcal{X} \times_m \boldsymbol{U})_{i_1 \ldots i_{m-1} j i_{m+1} \ldots i_M} = \sum_{i_m=1}^{I_m} x_{i_1 i_2 \ldots i_M} u_{j i_m}. \tag{1}$$

The mode-m (vector) product of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_M}$ with a vector $\boldsymbol{u} \in \mathbb{R}^{I_m}$ is denoted by $\mathcal{X} \times_m \boldsymbol{u} \in \mathbb{R}^{I_1 \times \cdots \times I_{m-1} \times I_{m+1} \times \cdots \times I_M}$. The result is of order $M-1$ and is defined element-wise as

$$(\mathcal{X} \times_m \boldsymbol{u})_{i_1 \ldots i_{m-1} i_{m+1} \ldots i_M} = \sum_{i_m=1}^{I_m} x_{i_1 i_2 \ldots i_M} u_{i_m}. \tag{2}$$

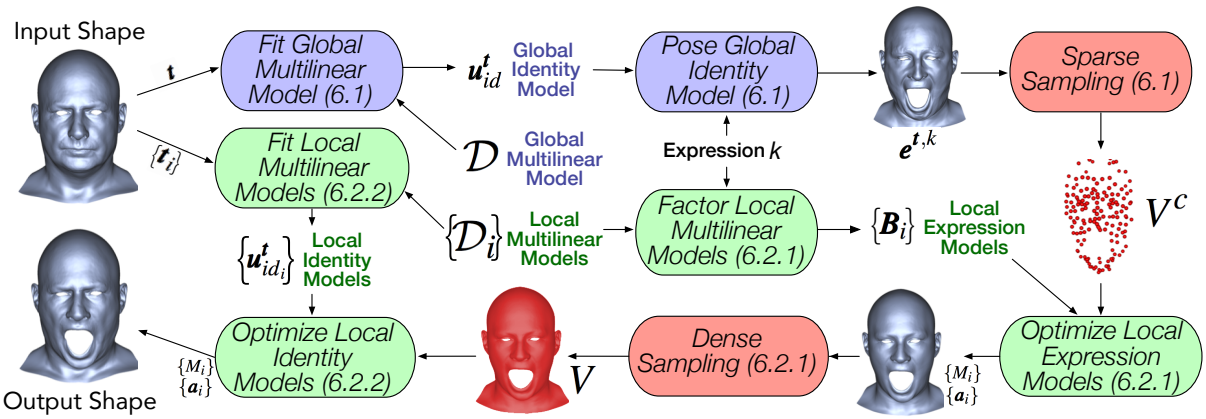

Figure 3: *Our approach uses a global multilinear model (blue) to estimate coarse deformation of the new expression. A sparse sampling of the global model result is used to then optimize the local models (green). The local expression models are optimized first to return a plausible new expression. The resulting shape is densely sampled to provide constraints when optimizing the local identity models which predict the identity of the input shape. The result is a high-quality, plausible new facial expression of the input shape. Actions such as model fitting are represented in beveled boxes.*

5. Data Acquisition

Our method for expression synthesis is data-driven, and thus we require a dataset of 3D facial expressions across different identities. Our approach is generic and can be applied to any 3D expression dataset with consistent mesh topology and common expressions across subjects. In this work we will demonstrate results on two different datasets. The first is 4DFAB [CKPZ18], which consists of 110 subjects performing 6 expressions at a resolution of 53K vertices. In order to provide richer expression results, we also propose our own dataset of 56 individuals performing 24 different expressions captured at a resolution of 95K vertices. Our dataset is built using the high-resolution facial scanner of Beeler et al. [BBB*10], where the face scans are put into vertex correspondence using a template mesh that is deformed through Laplacian deformation [SCOL*04] to exactly match the face scans, initialized by manual landmark annotations.

To remain general, let us define our data to be a set of 3D meshes with n_v vertices, spanning n_{id} subjects and undergoing n_{exp} expressions, represented as the tensor $\mathcal{D} \in \mathbb{R}^{n_v \times n_{id} \times n_{exp}}$. Throughout the method description we will use our new dataset as an example, and then show results on the 4DFAB dataset in Section 7.

6. Global-Local Expression Synthesis

We now present our global-local multilinear model for expression synthesis and discuss how the models are used to generate novel expressions given a target subject.

6.1. Global Model

We start by building a traditional multilinear model [VT02] using our training data $\mathcal{D} \in \mathbb{R}^{n_v \times n_{id} \times n_{exp}}$.

$$\mathcal{D} = \mathcal{C} \times_1 \boldsymbol{U}_{vertices} \times_2 \boldsymbol{U}_{identities} \times_3 \boldsymbol{U}_{expressions}, \tag{3}$$

where the core tensor $\mathcal{C} \in \mathbb{R}^{n_v \times n_{id} \times n_{exp}}$ governs the interactions between the different factors. The $n_{id} \times n_{id}$ mode matrix $\boldsymbol{U}_{identities}$ spans the space of people parameters, whereas the $n_{exp} \times n_{exp}$ mode matrix $\boldsymbol{U}_{expressions}$ spans the space of expression parameters. The $n_v \times n_v$ mode matrix $\boldsymbol{U}_{vertices}$ spans the space of 3D meshes. As the model is built from the entire face mesh, we call this our *global multilinear model*.

6.1.0.1. Global Model Fitting. The global model can be fit to a target neutral mesh $\boldsymbol{t}$ by selecting the row from $\boldsymbol{U}_{expressions}$ that corresponds to the neutral expression coefficients, denoted $\boldsymbol{u}_{exp}^{neutral}$, and then estimating the unknown identity coefficients $\boldsymbol{u}_{id}^{\boldsymbol{t}}$ of $\boldsymbol{t}$ within the model as

$$\boldsymbol{t} = (\mathcal{C} \times_1 \boldsymbol{U}_{vertices} \times_3 \boldsymbol{u}_{exp}^{neutral}) \times_2 \boldsymbol{u}_{id}^{\boldsymbol{t}}, \tag{4}$$

where $\boldsymbol{u}_{id}^{\boldsymbol{t}}$ is solved in a least-squares sense.

6.1.0.2. Global Model Posing. Once the identity coefficients are known, we can then predict the set of expressions $\boldsymbol{E}^{\boldsymbol{t}} \in \mathbb{R}^{n_v \times n_{exp}}$ corresponding to $\boldsymbol{t}$ using the global model by solving

$$\boldsymbol{E}^{\boldsymbol{t}} = \mathcal{C} \times_1 \boldsymbol{U}_{vertices} \times_2 \boldsymbol{u}_{id}^{\boldsymbol{t}} \times_3 \boldsymbol{U}_{expressions}, \tag{5}$$

or for a particular expression k

$$\boldsymbol{e}^{\boldsymbol{t},k} = \mathcal{C} \times_1 \boldsymbol{U}_{vertices} \times_2 \boldsymbol{u}_{id}^{\boldsymbol{t}} \times_3 \boldsymbol{u}_{exp}^{k}. \tag{6}$$

6.1.0.3. Sampling Coarse Deformation. The global model can represent important coarse deformation of the face when posing into expressions, but the local details tend to be inaccurate (see Figure 5 and the discussion in Section 6.3). Thus we wish to extract only the coarse expression deformation, which we do by sampling the surface at a sparse set of 163 locations, inspired by the

commonly-used paradigm of marker-based motion capture. Specifically, for a given expression k we select a subset of J vertices on the input mesh $\boldsymbol{t}$: $V^{\boldsymbol{t}} = \{\boldsymbol{v}_1^{\boldsymbol{t}}, ..., \boldsymbol{v}_J^{\boldsymbol{t}}\}$, and the corresponding vertices on the neutral expression $\boldsymbol{e}^{\boldsymbol{t},n}$ as well as $\boldsymbol{e}^{\boldsymbol{t},k}$, denoted V^n and V^k respectively. The corresponding vertices are selected directly from the 3D mesh. We then compute the sparse motion deltas induced by expression k and add them to the input mesh $\boldsymbol{t}$ to obtain sparse vertex samples V^c, as

$$\boldsymbol{v}_j^c = \boldsymbol{v}_j^{\boldsymbol{t}} + (\boldsymbol{v}_j^k - \boldsymbol{v}_j^n), \tag{7}$$

for each $\boldsymbol{v}_j^c \in V^c$ and corresponding $\boldsymbol{v}_j^{\boldsymbol{t}} \in V^{\boldsymbol{t}}$, $\boldsymbol{v}_j^k \in V^k$ and $\boldsymbol{v}_j^n \in V^n$. These sparse vertex positions encoding the coarse expression deformation are then passed to our local model fitting stages.

6.2. Local Models

A key component of our approach is to obtain the finer scale expression details using local multilinear models, built on subsets of the mesh vertices, distributed spatially across the face. We refer to a local subset of vertices as a *patch*. Figure 4 illustrates a semantically-meaningful artist-generated patch layout that we used (although our method can be applied with any layout), and note that patches are increased in size by 20% to provide sufficient overlap between neighboring patches for retaining smoothness in the final reconstruction. For each patch i, a local multilinear model is created from the data tensor $\mathcal{D}_i \in \mathbb{R}^{n_{v_i} \times n_{id} \times n_{exp}}$, where n_{v_i} is the number of vertices in patch i. This is analogous to the global model,

$$\mathcal{D}_i = \mathcal{C}_i \times_1 \boldsymbol{U}_{vertices_i} \times_2 \boldsymbol{U}_{identities_i} \times_3 \boldsymbol{U}_{expressions_i}, \tag{8}$$

however it is important to note that the patches in $\mathcal{D}_i$ are first rigidly aligned using Procrustes Analysis to the mean patch shape $\boldsymbol{m}_i$ computed over all identities and expressions, thus completely removing rigid patch motion and retaining only local deformations like stretching, compression, bulging, etc. The local multilinear models are thus not subsets of the global model. The individual local multilinear models each contain expression and identity modes, as illustrated in Figure 4.

Patch i can then be posed into a shape $\boldsymbol{x}_i$ using the patch reconstruction model (inspired by [WBGB16])

$$\boldsymbol{x}_i = M_i(\boldsymbol{m}_i + \sum_{f=1}^{F} \alpha_i^f \boldsymbol{b}_i^f), \tag{9}$$

where M_i is the rigid motion of the patch, $\{\boldsymbol{b}_i^1, ..., \boldsymbol{b}_i^F\} = \boldsymbol{B}_i$ is a deformation subspace for the patch consisting of F components, and $\{\alpha_i^1, ..., \alpha_i^F\}$ are the coefficients of the deformation basis. In terms of the deformation basis, we purposely choose to decouple identity and expression and solve for each one independently so as not to enforce explicit correlation between local identity and expression deformation. To accomplish this we can define the deformation subspace in two different ways, where $\boldsymbol{B}_i$ is composed of different slices of the local multilinear model, as described next.

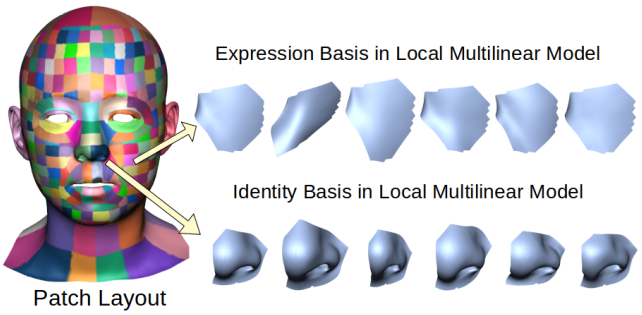

Figure 4: *We construct a local multilinear model for each patch in the given patch layout. The patch layout has been defined by an artist in a UV map. Each patch has its own expression basis and identity basis. A patch containing the nasolabial fold is visualized with its mean shape and top 5 PCA expression basis. We also visualize a nose patch in terms of its mean and top 5 PCA identity basis. Basis vectors are shown at* $+3\sigma$.

6.2.1. Local Expression Model

Given a particular expression k, we factor the local multilinear models to obtain a set of linear models, which we refer to as local expression models. The deformation space of these models spans the set of identities. Specifically, for each patch i we set

$$\boldsymbol{B}_i = \mathcal{C}_i \times_1 \boldsymbol{U}_{vertices_i} \times_2 \boldsymbol{U}_{identities_i} \times_3 \boldsymbol{u}_{exp_i}^k. \tag{10}$$

6.2.1.1. Optimizing Local Expression Models. The main idea is that we can use the sparse vertices V^c computed from the global model as constraints and then we can optimize the local expression models to obtain per-patch model parameters M_i and $\{\alpha_i^1, ..., \alpha_i^{n_{id}}\}$. Finally, we combine the patches to obtain a global face mesh with the desired expression. The details of the optimization procedure and mesh reconstruction are given in Section 6.3.

6.2.1.2. Sampling Dense Expression Deformation. The resulting mesh contains the desired local expression details with the desired coarse deformation, however it will not retain the local identity details of the target subject $\boldsymbol{t}$. Therefore, we proceed to fit and solve a local identity model as described next, however this time densely sampling the estimated expression mesh to create a dense set of vertex constraints, which allows to preserve the local expression deformations as closely as possible while solving for the identity.

6.2.2. Local Identity Model

We obtain the identity of the target subject by building a set of local identity models, one per patch, analogous to the local expression models but this time spanning the space of expressions for the desired identity.

6.2.2.1. Local Model Fitting. This is accomplished by solving for the local identity coefficients that best fit the local shape $\boldsymbol{t}_i$ of $\boldsymbol{t}$

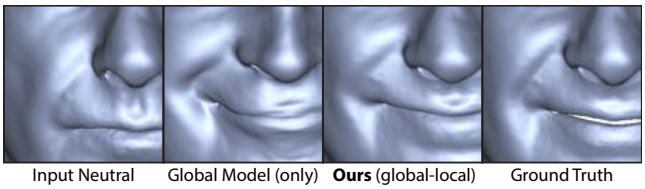

Figure 5: ***Local Details -*** *Given the input neutral shape, the global multilinear model contains artifacts when predicting a smile expression. Note the local details of the nose are incorrect, and the nasolabial fold is very different from the ground truth expression. Our global-local model achieves much more plausible results.*

corresponding to patch i. Specifically, we select the neutral expression coefficients $\boldsymbol{u}_{exp_i}^{neutral}$ from $\boldsymbol{U}_{expressions_i}$ and solve

$$\boldsymbol{t}_i = (\mathcal{C}_i \times_1 \boldsymbol{U}_{vertices_i} \times_3 \boldsymbol{u}_{exp_i}^{neutral}) \times_2 \boldsymbol{u}_{id_i}^{\boldsymbol{t}}, \tag{11}$$

for $\boldsymbol{u}_{id_i}^{\boldsymbol{t}}$ in a least-squares sense. Once obtaining $\boldsymbol{u}_{id_i}^{\boldsymbol{t}}$, we can then predict the expression patch shapes, which become the new deformation subspace $\boldsymbol{B}_i$

$$\boldsymbol{B}_i = \mathcal{C}_i \times_1 \boldsymbol{U}_{vertices_i} \times_2 \boldsymbol{u}_{id_i}^{\boldsymbol{t}} \times_3 \boldsymbol{U}_{expressions_i}. \tag{12}$$

6.2.2.2. Optimizing Local Identity Models. Using the dense vertex positions from the local expression solve as constraints, we optimize the local identity models to obtain final per-patch model parameters M_i and $\{\alpha_i^1, ..., \alpha_i^{n_{exp}}\}$, this time corresponding to the local identity models, and combine the patches to a global face mesh that forms the final synthesized expression k for target $\boldsymbol{t}$ (again, refer to Section 6.3 for optimization and mesh reconstruction details).

6.3. Local Model Optimization and Reconstruction

In the problem formulation for both the local expression model and the local identity model above, we must solve for the model parameters including the rigid local patch motion $\{M_i\}_{i=1}^{n_p}$, and the local blend coefficients $\{\boldsymbol{a}_i\}_{i=1}^{n_p}$ for the n_p patches. We denote $\boldsymbol{x}$ as our result mesh. We formulate the solution as an energy minimization problem.

$$\underset{\{M_i\}_{i=1}^{n_p}, \{\boldsymbol{a}_i\}_{i=1}^{n_p}}{\text{minimize}} \quad E_P + E_O + E_C, \tag{13}$$

where E_p is the *position* constraint (either the sparse or dense vertex positions defined above), E_o is the *overlap* constraint and E_c is the *subspace consistency* constraint, as described next.

6.3.1. Position Constraint

For the constrained subset of vertices V^c, obtained e.g. from the sparse sampling of the coarse deformation or the dense sampling of the expression deformation Equation 7, we formulate the position constraint as:

$$E_P = \sum_{\boldsymbol{v}_j^c \in V^c} \sum_{i \in \Omega(\boldsymbol{v}_j^c)} \|\boldsymbol{v}_j^c - \boldsymbol{x}_i(\boldsymbol{v}_j^c)\|^2, \tag{14}$$

where $\boldsymbol{v}_j^c$ denotes the positional constraint, $\Omega(\boldsymbol{v}_j^c)$ is the set of patches which contain vertex $\boldsymbol{v}_j^c$ and $\boldsymbol{x}_i(\boldsymbol{v}_j^c)$ refers to the corresponding vertex position of $\boldsymbol{v}_j^c$ within patch $\boldsymbol{x}_i$ from Equation 9.

6.3.2. Overlap Constraint

We employ an overlap constraint as defined in [WBGB16], which functions as a spatial regularizer as it encourages neighbouring patches to take on similar shapes in the overlapping area, defined as:

$$E_O = \lambda_O \sum_{\boldsymbol{v} \in S} \sum_{(i,j) \in \Omega(\boldsymbol{v}), i>j} \|\boldsymbol{x}_i(\boldsymbol{v}) - \boldsymbol{x}_j(\boldsymbol{v})\|^2, \tag{15}$$

where S is the set of vertices shared by patches, $\Omega(\boldsymbol{v})$ is the set of patches that contain vertex $\boldsymbol{v}$, $\boldsymbol{x}_i(\boldsymbol{v})$ is the 3D position of vertex $\boldsymbol{v}$ in patch i and $\lambda_O (= 0.1)$ is a weighting factor.

6.3.3. Subspace Consistency Constraint

In addition to the overlap constraint, we also add a subspace consistency term for neighbouring patches, incentivizing them to take on similar coefficients of deformation.

$$E_C = \lambda_C \sum_{\boldsymbol{v} \in S} \sum_{(i,j) \in \Omega(\boldsymbol{v}), i>j} \|\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_j\|^2, \tag{16}$$

where S and $\Omega(\boldsymbol{v})$ are the same as the overlap constraint, and $\boldsymbol{\alpha}_i$ corresponds to the deformation coefficients $\{\alpha_i^f\}_{f=1}^F$ of patch i and $\lambda_C (= 0.3)$ is a weighting factor.

6.3.4. Mesh Reconstruction

Once the model parameters are solved, all individual patches can be posed to form a global face shape, and we follow the approach of [WBGB16] to integrate the patches into a coherent shape.

As illustrated in Figure 5, our global-local synthesis method is able to produce plausible facial expressions with realistic local details that approximate the input identity, unlike the global multilinear model, which exhibits artifacts and does not retain the identity as well.

7. Results and Experiments

7.1. Qualitative Evaluation

We compare the results of our method against two baselines. One is the global multilinear model [VT02] commonly used in end-user facial animation applications, and the other is deformation transfer [SP04], often used in retargeting character animation in film

Figure 6: *We show the results of our method on [CKPZ18] and compare it against the baselines [VT02, SP04]. We used the result of [VT02] as the target expression for [SP04].*

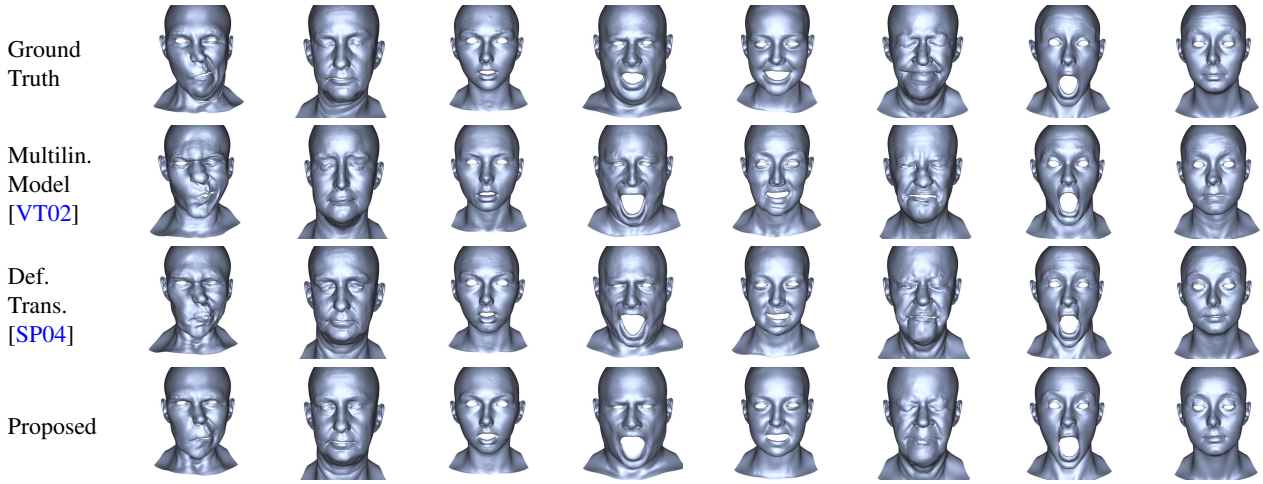

Figure 7: *We show the results of our method and compare it against the baselines [VT02, SP04]. We used the result of [VT02] as the target expression for [SP04].*

productions. For deformation transfer, we use the results of [VT02] as reference expression to transfer the deformations from.

We show the results of our method on two datasets: our own dataset containing high-quality facial meshes of 56 subjects under 24 expressions at a resolution of 95k vertices, and 4DFAB [CKPZ18], a dataset of 110 subjects under 6 expressions with meshes of 53k vertices. In the 4DFAB dataset, there is no neutral expression so we use the anger expression as our input meshes. We train our method on 100 subjects and test it on the remaining 10 subjects. Figure 6 shows how our method compares against the baselines [VT02, SP04] on 4DFAB. Our results are noticeably less noisy, approximate the identity well and exhibit plausible expressions. Figure 7 shows how our proposed method outperforms the baselines [VT02, SP04] on our own dataset, where we used 50 subjects for training and 6 for testing.

Another measure of qualitative evaluation is to render the synthesized expressions with the color texture of the neutral face, showing the alignment of geometric and texture features after synthesis. Figure 8 illustrates the textured results for a subset of synthesized expressions, showing that our method creates meaningful expres-

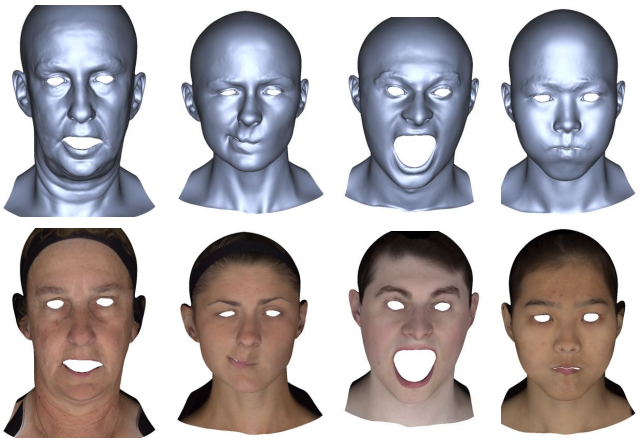

Figure 8: *Synthesized facial expressions (top row) rendered with neutral face texture (bottom row) show that our method creates meaningful expressions that are in vertex-correspondence with the neutral face.*

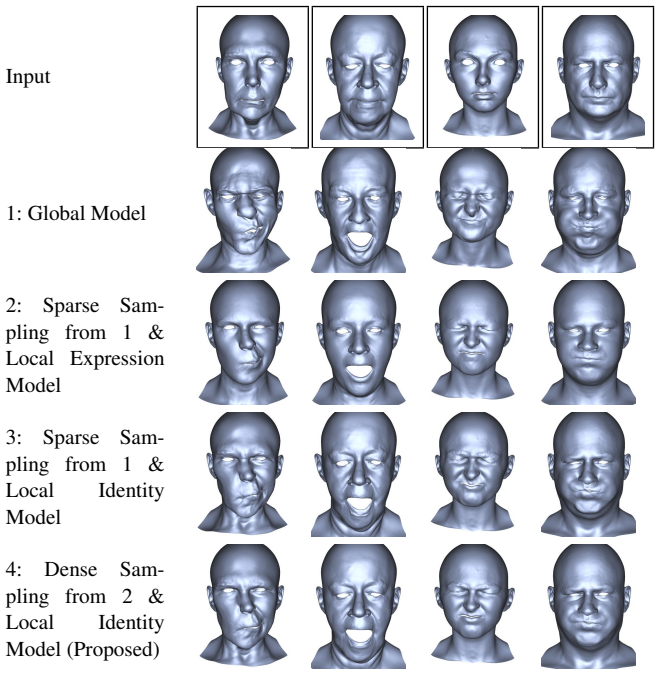

Figure 9: *Ablation Study: Our proposed method returns plausible expressions with the minimum amount of artifacts while maintaining the identity.*

sions that are in good vertex-correspondence with the input neutral face.

7.2. Ablation Study

We performed an ablation study to assess the impact of the different parts of our framework, as shown in Figure 9. For 4 different subjects, each with a different target expression to synthesize, we illustrate the results of 1 - purely global model fit, 2 - sparse sampling from 1 with a local expression model fit, 3 - sparse sampling from 1 with a local identity model fit (which contains artifacts), and our proposed method 4 - dense sampling from 2 with a local identity model fit. Method 3 reflects the result of applying local multilinear methods for facial expression synthesis. Though the details are preserved, the result expression is not plausible and contains artifacts. The proposed framework returns plausible expressions with minimal artifacts while maintaining the identity of the target subject.

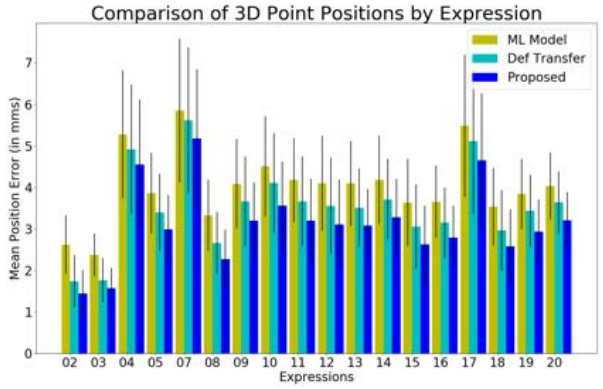

Figure 10: *Comparison grouped by expression of our proposed method with multilinear model [VT02] and deformation transfer [SP04] in terms of 3D position difference to the ground truth shape.*

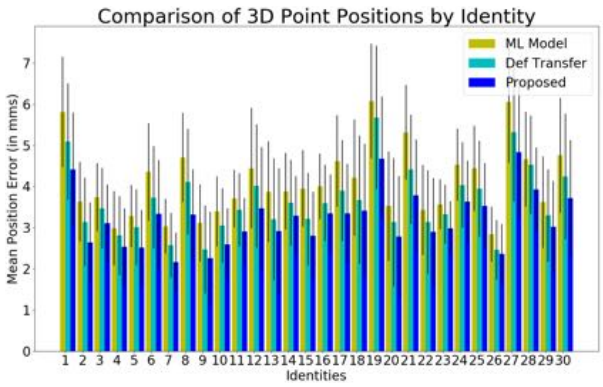

Figure 11: *Comparison grouped by identity of our proposed method with multilinear model [VT02] and deformation transfer [SP04] in terms of 3D position difference to the ground truth shape.*

7.3. Quantitative Evaluation

In order to quantitatively compare to existing methods, we predict 19 different expressions for 30 new test identities. We have removed high-level expressions such as sadness, happiness, and anger from the comparison since these are performed highly subjective and are hence not suited for quantitative analysis. For the 540 test meshes, we have ground truth results from the

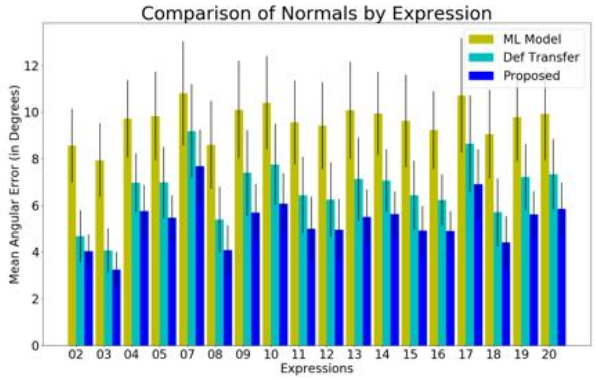

Figure 12: *Comparison grouped by expression of our proposed method with multilinear model [VT02] and deformation transfer [SP04] in terms of angular error between the normals of the predicted mesh and the normals of the ground truth shape.*

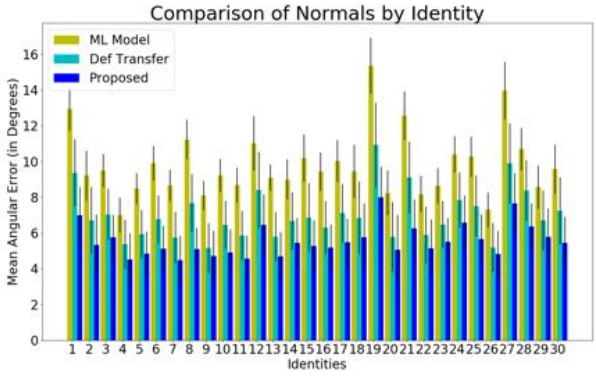

Figure 13: *Comparison grouped by identity of our proposed method with multilinear model [VT02] and deformation transfer [SP04] in terms of angular error between the normals of the predicted mesh and the normals of the ground truth shape.*

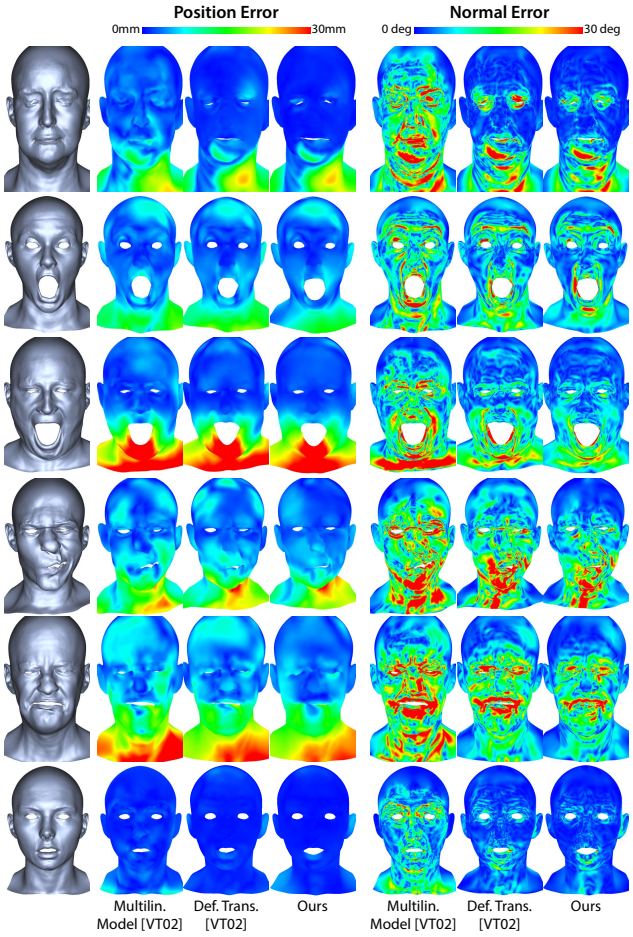

Figure 14: *Heatmap visualizations of position and normal errors for a selection of expressions in our dataset.*

dataset. We compare our proposed methods against the multilinear model [VT02] and deformation transfer [SP04] on 2 different metrics: 3D point position error in Figures 10 and 11 and angular error of the normals in Figures 12 and 13. By presenting the error values grouped by expression as well as by identity, we show that our method outperforms the baselines both in approximating the identity and in synthesizing the correct expression. Table 1 summarizes the comparison and Figure 14 visualizes a selection of expressions from our dataset with heatmaps showing the distribution of position and normal errors. We observe that our proposed method outperforms the baselines in both metrics.

Metric		[VT02]	[SP04]	Ours
Position error on our data	[mm]	4.03	3.54	**3.12**
Position error on [CKPZ18]	[mm]	2.32	2.23	**1.98**
Normal error on our data	[°]	9.63	6.72	**5.32**
Normal error on [CKPZ18]	[°]	6.01	5.75	**4.75**

Table 1: *Quantitative comparison reporting the mean position and mean normal angular error.*

7.4. Perceptual User Study

We conducted an anonymous user study where each user was presented with an image depicting a neutral expression of a random subject alongside the expressions generated by our proposed method and the two baseline methods [VT02, SP04]. Then the user was asked to choose which among the three options is the most plausible expression for the person in the neutral image. Each user was asked to rate a total of 50 samples spanning 24 different expressions and 25 different identities. The results are reported in Table 2. Our proposed method was chosen 57.60% of the time, more than twice as often than any of the baseline methods. A total of 50 users participated in the user study, yielding 2429 selections. The results are statistically significant with $p < 10^{-4}$.

Method	Top Selection Rate
Ours	**57.60%**
[SP04]	20.95%
[VT02]	21.45%

Table 2: *Users tend to select shapes synthesized by the proposed method as their preferred result twice as likely than shapes synthesized by the baselines.*

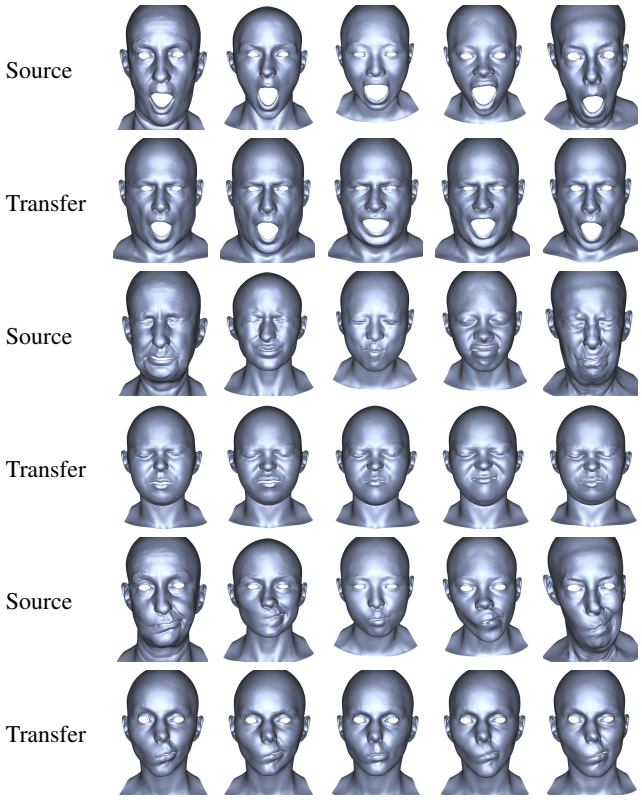

Figure 15: *Extracting the coarse deformation from individual subjects instead of the global model allows to synthesize a wide range of expression nuances while approximating the target identity, which is very valuable for data augmentation purposes.*

7.5. Data Augmentation

In the context of publicly available 3D face datasets, our expression synthesis method can be used for data augmentation purposes. Many available 3D face datasets contain only faces in neutral expression. By registering the template mesh of the new dataset to the template mesh of our dataset, the trained global-local model can be applied directly to meshes from a new dataset. Even more variability may be achieved by bypassing the global model estimate and directly transferring the coarse motion from the individual subjects instead, allowing to synthesize a wide range of expression nuances as shown in Figure 15, while still being plausible for the target identity.

8. Discussions and Future Work

Our method achieves realistic facial expression synthesis results despite limited training data. A weakness of our approach, however, is the necessity of a complete data tensor as training data. We require facial meshes of various people, all with the same set of expressions. It could be possible to replace the supervised multilinear model by an unsupervised one such as proposed by [WPSZ18] to leverage more "in-the-wild" datasets. Another future direction could be to utilize recent advances in geometric deep learning to replace multilinear models - this would require large 3D facial datasets.

9. Conclusion

We present a method for synthesizing realistic facial expressions given a target neutral face mesh. Leveraging a new global-local multilinear model, our method produces novel expressions that approximate the target identity, contain plausible expression details at various scales, and are free from geometric artifacts. Our method both quantitatively and qualitatively outperforms current state-of-the-art expression synthesis algorithms, and can be used to generate high-quality facial rigs or augment existing multi-identity facial datasets.

References

[BBB*10] BEELER T., BICKEL B., BEARDSLEY P., SUMNER B., GROSS M.: High-quality single-shot capture of facial geometry. *ACM Transactions on Graphics (TOG) 29*, 3 (2010), 40:1–40:9. 4

[BBPV03] BLANZ V., BASSO C., POGGIO T., VETTER T.: Reanimating faces in images and video. In *Computer graphics forum* (2003), vol. 22, Wiley Online Library, pp. 641–650. 2

[BBW14] BRUNTON A., BOLKART T., WUHRER S.: Multilinear wavelets: A statistical shape space for human faces. In *Computer Vision – ECCV 2014* (Cham, 2014), Fleet D., Pajdla T., Schiele B., Tuytelaars T., (Eds.), Springer International Publishing, pp. 297–312. 2

[BRZ*16] BOOTH J., ROUSSOS A., ZAFEIRIOU S., PONNIAH A., DUNAWAY D.: A 3d morphable model learnt from 10,000 faces. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (6 2016). 2, 3

[BV*99] BLANZ V., VETTER T., ET AL.: A morphable model for the synthesis of 3d faces. In *Siggraph* (1999), vol. 99, pp. 187–194. 1, 2

[CCK*18] CHOI Y., CHOI M., KIM M., HA J.-W., KIM S., CHOO J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 8789–8797. 2

[CKPZ18] CHENG S., KOTSIA I., PANTIC M., ZAFEIRIOU S.: 4dfab: A large scale 4d database for facial expression analysis and biometric applications. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018)* (Salt Lake City, Utah, US, 6 2018). 3, 4, 7, 9

[CWZ*14] CAO C., WENG Y., ZHOU S., TONG Y., ZHOU K.: Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics 20*, 3 (2014), 413–425. 2

[DN08] DENG Z., NOH J.: Computer facial animation: A survey. In *Data-driven 3D facial animation*. Springer, 2008, pp. 1–28. 1

[LAR*14] LEWIS J. P., ANJYO K., RHEE T., ZHANG M., PIGHIN F. H., DENG Z.: Practice and theory of blendshape facial models. *Eurographics (State of the Art Reports) 1*, 8 (2014), 2. 1

[LBB*17] LI T., BOLKART T., BLACK M. J., LI H., ROMERO J.: Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics (TOG) 36*, 6 (2017), 194. 2

[PAM*18] PUMAROLA A., AGUDO A., MARTINEZ A. M., SANFELIU A., MORENO-NOGUER F.: Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 818–833. 2

[SCOL*04] SORKINE O., COHEN-OR D., LIPMAN Y., ALEXA M., RÖSSL C., SEIDEL H.-P.: Laplacian surface editing. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing* (2004), ACM, pp. 175–184. 4

[SP04] SUMNER R. W., POPOVIĆ J.: Deformation transfer for triangle meshes. *ACM Trans. Graph. 23*, 3 (Aug. 2004), 399–405. 2, 6, 7, 8, 9, 10

[SYH*17] SHU Z., YUMER E., HADAP S., SUNKAVALLI K., SHECHTMAN E., SAMARAS D.: Neural face editing with intrinsic image disentangling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 5541–5550. 2

[VBPP05] VLASIC D., BRAND M., PFISTER H., POPOVIĆ J.: Face transfer with multilinear models. *ACM Trans. Graph. 24*, 3 (July 2005), 426–433. 2, 3

[VT02] VASILESCU M. A. O., TERZOPOULOS D.: Multilinear analysis of image ensembles: Tensorfaces. In *Computer Vision — ECCV 2002* (Berlin, Heidelberg, 2002), Heyden A., Sparr G., Nielsen M., Johansen P., (Eds.), Springer Berlin Heidelberg, pp. 447–460. 2, 3, 4, 6, 7, 8, 9, 10

[WBGB16] WU C., BRADLEY D., GROSS M., BEELER T.: An anatomically-constrained local deformation model for monocular face capture. *ACM Trans. Graph. 35*, 4 (July 2016), 115:1–115:12. 5, 6

[WPSZ18] WANG M., PANAGAKIS Y., SNAPE P., ZAFEIRIOU S. P.: Disentangling the modes of variation in unlabelled data. *IEEE Transactions on Pattern Analysis and Machine Intelligence 40*, 11 (11 2018), 2682–2695. 2, 10

[WSC*19] WANG M., SHU Z., CHENG S., PANAGAKIS Y., SAMARAS D., ZAFEIRIOU S.: An adversarial neuro-tensorial approach for learning disentangled representations. *International Journal of Computer Vision* (6 2019). 2

[YCS*08] YIN L., CHEN X., SUN Y., WORM T., REALE M.: A high-resolution 3d dynamic facial expression database. In *2008 8th IEEE International Conference on Automatic Face Gesture Recognition* (9 2008), pp. 1–6. 3

[YWS*06] YIN L., WEI X., SUN Y., WANG J., ROSATO M. J.: A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)* (2006), IEEE, pp. 211–216. 3

[ZLZ*14] ZHANG Y., LIN W., ZHOU B., CHEN Z., SHENG B., WU J.: Facial expression cloning with elastic and muscle models. *Journal of Visual Communication and Image Representation 25*, 5 (2014), 916–927. 3

[ZTG*18] ZOLLHÖFER M., THIES J., GARRIDO P., BRADLEY D., BEELER T., PÉREZ P., STAMMINGER M., NIESSNER M., THEOBALT C.: State of the art on monocular 3d face reconstruction, tracking, and applications. In *Computer Graphics Forum* (2018), vol. 37, Wiley Online Library, pp. 523–550. 1