International Journal of Computer Vision (2021) 129:2680–2713
https://doi.org/10.1007/s11263-021-01494-4

Shape My Face: Registering 3D Face Scans by Surface-to-Surface Translation

Mehdi Bahri[1] · Eimear O' Sullivan[1] · Shunwang Gong[1] · Feng Liu[2] · Xiaoming Liu[2] · Michael M. Bronstein[1] · Stefanos Zafeiriou[1]

Received: 15 September 2020 / Accepted: 18 June 2021 / Published online: 10 July 2021

Abstract

Standard registration algorithms need to be independently applied to each surface to register, following careful pre-processing and hand-tuning. Recently, learning-based approaches have emerged that reduce the registration of new scans to running inference with a previously-trained model. The potential benefits are multifold: inference is typically orders of magnitude faster than solving a new instance of a difficult optimization problem, deep learning models can be made robust to noise and corruption, and the trained model may be re-used for other tasks, e.g. through transfer learning. In this paper, we cast the registration task as a surface-to-surface translation problem, and design a model to reliably capture the latent geometric information directly from raw 3D face scans. We introduce Shape-My-Face (SMF), a powerful encoder-decoder architecture based on an improved point cloud encoder, a novel visual attention mechanism, graph convolutional decoders with skip connections, and a specialized mouth model that we smoothly integrate with the mesh convolutions. Compared to the previous state-of-the-art learning algorithms for non-rigid registration of face scans, SMF only requires the raw data to be rigidly aligned (with scaling) with a pre-defined face template. Additionally, our model provides topologically-sound meshes with minimal supervision, offers faster training time, has orders of magnitude fewer trainable parameters, is more robust to noise, and can generalize to previously unseen datasets. We extensively evaluate the quality of our registrations on diverse data. We demonstrate the robustness and generalizability of our model with in-the-wild face scans across different modalities, sensor types, and resolutions. Finally, we show that, by learning to register scans, SMF produces a hybrid linear and non-linear morphable model. Manipulation of the latent space of SMF allows for shape generation, and morphing applications such as expression transfer in-the-wild. We train SMF on a dataset of human faces comprising 9 large-scale databases on commodity hardware.

Keywords Surface registration · Non linear morphable models · Face modeling · Point cloud · Graph neural network · Generative modeling

1 Introduction

3D shapes come in a variety of representations, including range images, voxel grids, point clouds, implicit surfaces, and meshes. Human face scans, in particular, are often given as either range images, or meshes, but typically do not share a common parameterization (i.e., the output of the 3D scanner

Communicated by Gregory Rogez.

✉ Mehdi Bahri
m.bahri@imperial.ac.uk

Eimear O' Sullivan
e.o-sullivan16@imperial.ac.uk

Shunwang Gong
shunwang.gong16@imperial.ac.uk

Feng Liu
isliuf1990@gmail.com

Xiaoming Liu
liuxm@cse.msu.edu

Michael M. Bronstein
m.bronstein@imperial.ac.uk

Stefanos Zafeiriou
s.zafeiriou@imperial.ac.uk

[1] Department of Computing, Imperial College London, London, UK

[2] Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA

does not typically have a fixed connectivity, sampling rate etc.). Fundamentally, this diversity of representations is only a by-product of the inability of computers to represent continuous surfaces, but the latent geometric information to be represented is the same. In practice, this poses a challenge: two surfaces represented with two different parameterizations are not easily compared, which makes exploiting the geometric information difficult. Finding a shared representation while preserving the geometry is the task of dense surface registration, a cornerstone in both 3D computer vision and graphics (Amberg et al. 2007; Salazar et al. 2014).

The design and construction of a shared shape representation is often implemented by means of a common template, which has a predefined number of vertices and vertex connectivity. After choosing the common template, a fitting method is implemented to bring the raw facial scans in dense correspondence with the chosen template. The use of a common template is a crucial step towards learning a statistical model of the face shape, also know as 3D Morphable Models (3DMMs) (Blanz and Vetter 1999; Booth et al. 2016), which is a very important tool for shape representation and has been used for a wide range of applications spanning from 3D face reconstruction from images (Blanz and Vetter 2003; Booth et al. 2018b) to diagnosis and treatment of face disorders (Knoops et al. 2019; Mueller et al. 2011).

Arguably, the current methods of choice for establishing dense correspondences are variants of Non-rigid Iterative Closest Point (NICP) (Amberg et al. 2007), and non-rigid registration approaches whose regularization properties are defined by statistical (Cheng et al. 2017) and non-statistical (Lüthi et al. 2018) models. The application of deep learning techniques to the problem of establishing dense correspondences was only recently possible after the design of proper layered structures that directly consumes point clouds and respect the permutation invariance of points in the input data (e.g., PointNet (Qi et al. 2017a)).

To the best of our knowledge the only technique that tries to solve the problem of establishing dense correspondences on unstructured point-cloud data and learning a face model on a common template has been presented in Liu et al. (2019). The method uses a PointNet to summarise (i.e., encode) the information of an unstructured facial point cloud. Then, fully-connected layers (similar to the ones used in dense statistical models (Blanz and Vetter 1999; Booth et al. 2016)) are used to reconstruct (i.e., decode) the geometric information in the topology of the common template. In this paper, we work on a similar line of research and we make a series of important contributions in three different areas. In particular,

- **Network architecture** We propose architectural modifications of the point cloud CNN framework that improve on restrictions of Qi et al. (2017a). That is, in order to avoid having to adopt heuristic noise reduction and cropping strategies we incorporate a learned attention mechanism in the network structure. We demonstrate that the proposed architecture is better suited for in-the-wild captured data. Furthermore, we propose a variant of PointNet better suited for small batches, hence able to consume higher resolution raw-scans. Our morphable model part of the network (i.e., the decoder) comprises of a series of mesh-convolutional layers (Bouritsas et al. 2019; Gong et al. 2019) with novel (in the mesh processing literature) skip connections that can capture better details and local structures. Finally, our network structure is also considerably smaller than the state-of-the-art.
- **Engineering/Implementation** One of the major challenges when establishing dense correspondences in raw facial scans is the large deformations of the mouth area, especially in extreme expressions. We propose a very carefully engineered approach that smoothly incorporates a statistical mouth model. We demonstrate our method captures the mouth area very robustly.
- **Application** Our emphasis in this work is on robustness to noise in the scans (e.g. sensor noise, background contamination, and points from the inside of the mouth), compactness of the model, and generalization. The model we develop should be readily usable on, e.g., embedded 3D scanners to produce both a registered scan and a set of latent representations that can be leveraged in downstream tasks. We present extensive experiments to demonstrate the power of our algorithm, such as expression transfer and interpolation between in the wild scans across modalities and resolution. One of the major outcomes of our paper is a novel morphable model trained on 9 diverse large scale datasets, which will be made public.

Figure 1 shows some test textured scans and their corresponding registrations and attention masks.

1.1 Structure of the Paper

We provide an extensive summary of prior published work in Sect. 2, covering relevant areas of the morphable models, registration, and 3D deep learning literature. Section 3 is dedicated to reviewing the current state of the art model, which we use as a baseline in our experiments, and to highlight the limitations and challenges we tackle. We introduce our model, Shape My Face (SMF) in Sect. 4, and provide detailed descriptions of its different components, how they provide solutions to the challenges identified in Sect. 3, and how they allow us to frame the registration task as a surface-to-surface translation problem. We also introduce our model trained on a very large dataset comprising 9 large human face scans databases. For the sake of clarity, we split our experimental evaluation into two parts. Section 5 studies the performance of SMF for registration, and presents a statis-

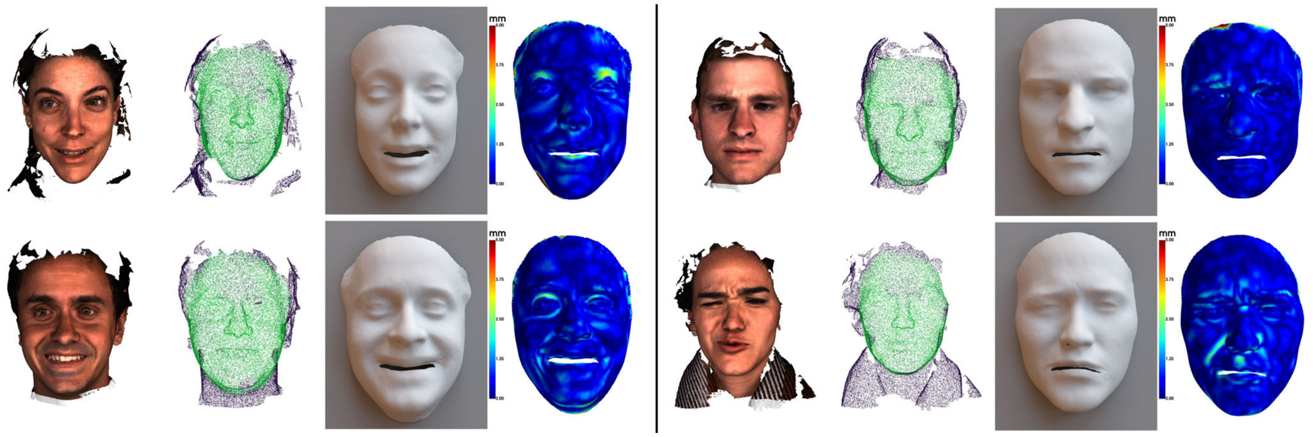


Fig. 1 Sample test scans and their registration. Left to right: textured mesh, input point cloud sampled uniformly from the mesh (black) and the attention mask predicted by the model (green), registration, and heatmap of the surface error

tical analysis of the model's stability, as well as an ablation study. Section 6 evaluates SMF on morphable model applications and studies properties of the latent representations; in particular, in Sect. 6.4 we evaluate SMF on surface-to-surface translation applications entirely in the wild.

Notations Throughout the paper, matrices and vectors are denoted by upper and lowercase bold letters (e.g., $\mathbf{X}$ and ($\mathbf{x}$), respectively. $\mathbf{I}$ denotes the identity matrix of compatible dimensions. The i^{th} column of $\mathbf{X}$ is denoted as $\mathbf{x}_i$. The sets of real numbers is denoted by $\mathbb{R}$. A *graph* $\mathscr{G} = (\mathscr{V}, \mathscr{E})$ consists of *vertices* $\mathscr{V} = \{1, \dots, n\}$ and *edges* $\mathscr{E} \subseteq \mathscr{V} \times \mathscr{V}$. The graph structure can be encoded in the *adjacency matrix* $\mathbf{A}$, where $a_{ij} = 1$ if $(i, j) \in \mathscr{E}$ (in which case i and j are said to be *adjacent*) and zero otherwise. The *degree matrix* $\mathbf{D}$ is a diagonal matrix with elements $d_{ii} = \sum_{j=1}^{n} a_{ij}$. The *neighborhood* of vertex i, denoted by $\mathscr{N}(i) = \{j : (i, j) \in \mathscr{E}\}$, is the set of vertices adjacent to i.

2 Related Work

Although primarily a fast registration method with a focus on generalizability to unseen data, our approach also makes important progress towards learning an accurate part-based non-linear 3D morphable model of the human face, as well as a generative model with applications to surface-to-surface translation. We first review the relevant literature across the related fields. Then, we devote Sect. 3 to exposing the limitations of the current state of the art algorithm that motivate the choices made in this work.

2.1 Surface Registration and Statistical Morphable Models

Surface registration is the task of finding a common parameterization for heterogeneous surfaces. It is a necessary pre-processing step for a range of downstream tasks that assume a consistent representation of the data, such as statistical analysis and building 3D morphable models. As such, it is a fundamental problem in 3D computer vision and graphics.

2.1.1 Surface Registration

Two main classes of methods coexist for surface registration. Image-based registration methods first require finding a mapping between the surface to align and a two-dimensional parameter space; most commonly, a UV parameterization is computed for a textured mesh, typically using a cylindrical projection. Image registration methods are then applied to align the unwrapped surface with a template, for instance using optical flow analysis (Horn and Schunck 1981; Lefébure and Cohen 2001), or thin plate spline warps (Bookstein 1989). UV-space registration is computationally efficient and relies on mature image processing techniques, but the flattening step unavoidably leads to a loss of information, and sampling of the UV space is required to reconstruct a surface. For this reason, the second main class of surface registration methods operates directly in 3D, avoiding the UV space entirely. Prominent examples include the Non Rigid Iterative Closest Point (NICP) method (Amberg et al. 2007), a generalization of the Iterative Closest Point (ICP) method (Chen and Medioni 1991; Besl and McKay 1992) that introduces local deformations, or the Coherent Point Drift (CPD) algorithm (Myronenko et al. 2007; Myronenko and Song 2010). NICP operates on meshes and solves a non-convex energy minimization problem that encourages the vertices of the registered mesh to be close to the target surface, and the local transformations to be similar for spatially close points. Due to its non-convex nature, NICP is sensitive to initialization, and is most often used in conjunction with sparse annotations (i.e. landmarks for which a 1-to-1 correspondence is known a priori). Similarly, CPD also encourages the motion of neighboring points to be similar, but operates on point clouds and frames the registration problem as that of mass matching between probability distributions. As such, it is closely related to optimal transport registration (Feydy et al. 2017). We refer to relevant surveys (van Kaick et al. 2011; Tam et al. 2013) for a more complete review of non-deep learning based surface registration methods.

2.1.2 Linear, Multilinear, and Non-linear Morphable Models

Linear morphable models for the human face were first introduced in the seminal work of Blanz and Vetter (1999). The authors proposed to model the variability of human facial anatomy by applying Principal Component Analysis (PCA) (Pearson 1901; Hotelling 1933) to 200 laser scans (100 male and 100 female) of young adults in a neutral pose. Scans were aligned by image registration in the UV space with a regularized form of optical flow. The resulting set of components forms an orthogonal basis of faces that can be manipulated to synthesize new faces. Amberg et al. (2008) extended the PCA approach to handle expressions for expression invariant 3D face recognition, using scans registered directly with NICP (Amberg et al. 2007). Patel and Smith (2009) introduced the widely-used Basel Face Model (BFM), also trained on 200 scans registered with NICP. It is only with the work of Booth et al. (2016, 2018a) that a morphable model trained on a large heterogeneous population, known as the Large Scale Face Model (LSFM) was made available. The authors use the BFM template and a modification of the NICP algorithm, along with automated pruning strategies, to build a high quality model of the human face from almost 10000 subjects. LSFM is trained on neutral scans only, but can be combined with a bank of facial expressions, such as the popular FaceWarehouse (Cao et al. 2014).

Multilinear extensions of linear morphable models have been considered as early as Vlasic et al. (2005) where a tensor factorization was used to model different modes of variation independently (e.g., identity and expression) with applications to face transfer, and refined by Bolkart and Wuhrer (2015). However, the multilinear approach requires every combination of subject and expression to be present exactly once in the dataset, a requirement that can be both hard to satisfy and limiting in practice. Salazar et al. (2014) proposed an explicit decomposition into blendshapes as an alternative. In Li et al. (2017), the authors propose to combine an articulated jaw with linear blending to obtain a non-linear model of facial expressions.

2.1.3 Part-Based Models

Besides a global PCA model, Blanz and Vetter (1999) also presented a part-based morphable model. The authors manually segmented the face into separate regions and trained specialized 3DMMs for each part, that can then be morphed independently. The resulting model is more expressive than a global PCA would be, and is obtained by combining the parts using a modification of the image blending algorithm of Burt and Adelson (1985). De Smet and Van Gool (2011) and Tena et al. (2011) showed manual segmentation may not be optimal, and that better segmentation can be defined by statistical analysis. Tena et al. (2011) designed an interpretable region-based model for facial animation purposes.

Part-based models also appear when attempting to represent together different distinct parts of the body. Romero et al. (2017) model hands and bodies together by replacing the hand region of SMPL (Loper et al. 2015) with a new specialized hand model called MANO. Joo et al. (2018) present the *Frankenstein* model, a morphable model of the whole human body that combines existing specialized models of the face (Cao et al. 2014), body (Loper et al. 2015), and a new artist-generated model for hands. The model's parameters are defined as the concatenation of all the parts' parameters. The final reconstruction is obtained by linear blending of the vertices of the separate parts using a manually-crafted matrix. The final model has fewer vertices than the sum of its parts, and the parts were manually aligned. As per the author's own description, minimal blending is done at the seams.

In Ploumpis et al. (2019, 2020), a high-definition head and face model is created by blending together the Liverpool-York Head model (LYHM) (Dai et al. 2017) and the Large-Scale Face Model (LSFM) (Booth et al. 2018a). While LYHM includes a facial region, replacing it with LSFM offers more details. Two approaches are proposed to combine the models smoothly. A regression model learned between the two models' parameter spaces, and a Gaussian Process Morphable Model (GPMM) approach (Luthi et al. 2018) where the covariance matrix of a GPMM is carefully crafted from the covariance matrices of its parts using a weighting scheme based on the Euclidean distance of the vertices to the nose tip of the registered meshes (i.e. the outputs of the head and face models). A refinement phase involving non-rigid ICP further tunes the covariance matrix of the GPMM.

We refer the interested reader to the recent review of Egger et al. (2020) for more information.

2.2 Deep Learning on Surfaces

Deep neural networks now permeate computer vision, but have only become prominent in 3D vision and graphics in the past few years. We review some of the recent algorithmic advances for representation learning on surfaces, surface registration, and morphable models.

2.2.1 Geometric Deep Learning on Point Clouds and Meshes

Recent methods from the field of Geometric Deep Learning (Bronstein et al. 2017) have emerged and propose analogues of classical deep learning operations such as convolutions for meshes and point clouds.

Point cloud processing methods treat the discrete surface as an unordered point set, with no pre-defined notion of intrinsic distances or connectivity. The pioneering work of PointNet (Qi et al. 2017a) defines a point set processing layer as a 1×1 convolution shared among all points, followed by batch normalization, and ReLU activation. The resulting local point-wise features are aggregated into a global representation of the surface by max pooling. In spite of its simplicity, PointNet achieved state of the art result in both 3D object classification and point cloud segmentation tasks, and remains competitive to this day. Follow-up works have explored extending PointNet to enable hierarchical feature learning (Qi et al. 2017b), as well as more powerful architectures that attempt to learn the metric of the surface via local kernel functions (Xu et al. 2018; Lei et al. 2019; Zhang et al. 2019), or by building a k-NN graph in the feature space (Wang et al. 2019). While these methods obtain higher classification and segmentation accuracy, their computational complexity limits their application to large-scale point clouds, a task for which PointNet is often preferred.

Graph Neural Networks, on the other hand, assume the input to be a graph, which naturally defines connectivity and distances between points. Initial formulations were based on the convolution theorem and defined graph convolutions using the graph Fourier transform, obtained by eigenanalysis of the combinatorial graph Laplacian (Bruna et al. 2014), and relied on smoothness in the spectral domain to enforce spatial locality. Defferrard et al. (2016) accelerated spectral graph CNNs by expanding the filters on the orthogonal basis of Chebyshev polynomials of the graph Laplacian, also providing naturally localized filters. However, the Laplacian is topology-specific which hurts the performance of these methods when a fixed connectivity cannot be guaranteed. Kipf and Welling (2017) further simplified graph convolutions by reducing ChebNet to its first order expansion, merging trainable parameters, and removing the reliance on the eigenvalues of the Laplacian. The resulting model, GCN, has been shown to be equivalent to Laplacian smoothing (Li et al. 2018) and has not been successful in shape processing applications. Attention-based models (Monti et al. 2017; Fey et al. 2018; Verma et al. 2018; Veličković et al. 2018) dynamically compute weighted features of a vertex's neighbours and do not expect a uniform connectivity in the dataset, and generalize the early spatial mesh CNNs that operated on pre-computed geodesic patches (Masci et al. 2015; Boscaini et al. 2016). Spatial and spectral approaches have both been shown to derive from the more general neural message passing (Gilmer et al. 2017) framework. Recently, SpiralNet (Lim et al. 2018), a specialized operator for meshes, has been introduced based on a consistent sequential enumeration of the neighbors around a vertex. Gong et al. (2019) introduces a refinement of the SpiralNet operator coined SpiralNet++ which simplifies the computation of the spiral patches.

Finally, recent work explored skip connections to help training deep graph neural networks. In Appendix B of Kipf and Welling (2017), the authors propose a residual architecture for deep GCNs. Hamilton et al. (2017) introduce an architecture for inductive learning on graphs based on an aggregation step followed by concatenation of the previous feature map and transformation by a fully-connected layer. Li et al. (2019) study very deep variants of the Dynamic Graph CNN (Wang et al. 2019) using residual and dense connections for point cloud processing. Finally, in Gong et al. (2020), the authors relate graph convolution operators to radial basis functions to propose affine skip connections, and demonstrate improved performance compared to vanilla residuals for a range of operators.

2.2.2 Registration

The methods presented in Sect. 2.1.1 are framed as optimization problems that need to be solved for every surface individually. Although able to produce highly accurate registrations, they can be costly to apply to large datasets, and are based on axiomatic conceptualizations of the registration task. The reliance on sparse annotations to accurately register expressive scans also means the data needs to be manually annotated, a tedious and expensive task. A new class of learning-based surface registration models is therefore emerging that, once passed the initial training effort, promise to reduce the registration of new data to a fast inference pass, and to potentially outperform hand-crafted algorithms. In PointNetLK (Aoki et al. 2019), the authors adapt the image registration of Lucas and Kanade (1981) to point clouds in a supervised learning setting. A PointNet (Qi et al. 2017a) encoder is trained to predict a rigid body transformation $\mathbf{G} \in SE(3)$, with a loss defined between the network's prediction $\mathbf{G}_{est}$ and a ground truth transformation $\mathbf{G}_{gt}$ as $||\mathbf{G}_{est}^{-1}\mathbf{G}_{gt} - \mathbf{I}||_F$, with $||.||_F$ the Frobenius (matrix ℓ_2) norm. A similar technique is employed in Wang and Solomon (2019a), where the authors introduce a supervised learning model for rigid registration coined as Deep Closest Point (DCP). DCP learns to predict the parameters of a rigid motion to align two point clouds, and is trained on synthetically generated pairs of point clouds, for which the ground truth parameters are known. The follow-up work of PRNet (Wang and Solomon 2019b) offers a self-supervised approach for learning rigid registration between partial point clouds. In Lu et al. (2019), and Li and Zhang (2019), supervised learning algorithms are defined for rigid registration, but with losses defined on dense correspondences between points, and on a soft-assigment matrix, respectively. Finally, Shimada et al. (2019) designed a U-Net like architecture on voxel grids for non-rigid point set registration, however, their method is limited by the resolution of the grid and does not build latent representations of the scans, nor does it provide a morphable model.

2.2.3 Morphable Models

Abrevaya et al. (2018) train a hybrid encoder-decoder architecture on rendered height maps from 3D face scans using an image CNN encoder and a multilinear decoder. This approach circumvents the need for prior registration of the scans to a template, but the face model itself remains linear.

Concurrently, there has been a surge of interest for deep non-linear morphable models to better capture extreme variations. Bagautdinov et al. (2018) model facial geometry in UV space with a variational auto-encoder (VAE). Tran and Liu (2018) replace the linear bases with fully-connected decoders to model 3D geometry and texture from images, a technique extended in Tran et al. (2019). Ranjan et al. (2018) introduce a convolutional mesh auto-encoder based on Chebyshev graph convolutions (Defferrard et al. 2016). Bouritsas et al. (2019), use Spiral Convolutions (Lim et al. 2018) to learn non-linear morphable models of bodies and faces. In both these works, the connectivity of the 3D meshes is assumed to be fixed; that is, the scans have to be registered a priori. The non-linear deep neural network replaces the PCA for dimensionality reduction.

In Liu et al. (2019), an asymetric autoencoder is proposed. A PointNet encoder is applied to rigidly aligned heterogeneous raw scans, and two fully-connected decoders produce identity and expression blendshapes independently on the BFM face template. Thus, the algorithm produces a registration of the input scan. Mesh convolutional decoders are proposed in Kolotouros et al. (2019b) for human body reconstruction from single images. In Kolotouros et al. (2019a), model-fitting is introduced to also produce representations directly on the SMPL model.

3 State of the Art

The autoencoder architecture of Liu et al. (2019) is the current state of the art for the learned registration of 3D face scans. A learning-based approach for registration is desirable since a model that generalizes would be able to register new scans very quickly, thus potentially offsetting the time spent training the model. Other benefits compared to traditional optimization-based registration may include increased robustness to noise in the data. Furthermore, an autoencoder learns an efficient latent representation of the scans, which may later be processed for other applications, while the trained decoder can be used in isolation as a morphable model.

Motivated by the aforementioned potential upsides, we review the approach of Liu et al. (2019) and identify key limitations and areas of improvement. We further evaluate a pre-trained model provided by the authors of Liu et al. (2019) on the same dataset used in the original paper (also provided by the authors). We refer to the provided pre-trained model as *the baseline*.

3.1 Problem Formulation and Architecture

A crop of the mean face of the BFM 2009 model is chosen as a face template on which to register the raw 3D face scans. A *registered* (densely aligned) face is modeled as an identity shape with an additive expression deformation:

$$\mathbf{S} = \mathbf{S}_{id} + \boldsymbol{\Delta}\mathbf{S}_{exp} \tag{1}$$

With $\mathbf{S} = [x_1, y_1, z_1; \ldots; x_N, y_N, z_N]$ the concatenated, consistently ordered, Cartesian 3D coordinates of the vertices. For this template, $N = 29495$.

A subset of N_s vertices from a processed input scan (details of the processing below) are sampled at random to obtain a point cloud representation of the scan. A vanilla PointNet encoder without spatial transformers produces a joint embedding $\mathbf{z}_{joint} \in \mathbb{R}^{1024}$. Two fully-connected (FC) layers, without non-linearities, are applied in parallel to obtain identity and expression latent vectors in $\mathbb{R}^{512}$:

$$\mathbf{z}_{id} = \mathbf{W}_{id} \cdot \mathbf{z}_{joint} + \mathbf{b}_{id} = \mathrm{FC}_{id}(\mathbf{z}_{joint}) \tag{2}$$

$$\mathbf{z}_{exp} = \mathbf{W}_{exp} \cdot \mathbf{z}_{joint} + \mathbf{b}_{exp} = \mathrm{FC}_{exp}(\mathbf{z}_{joint}). \tag{3}$$

Two multi-layer perceptrons consisting of two fully-connected layers with ReLU activations decode the identity and expression blendshapes from their corresponding vectors:

$$\begin{aligned} \mathbf{S}_{id} &= \mathrm{FC}^2_{id}\left(\xi\left(\mathrm{FC}^1_{id}(z_{id})\right)\right) & (4)\\ &= \mathrm{FC}^2_{id}\left(\xi\left(\mathrm{FC}^1_{id}\left(\mathrm{FC}_{id}(z_{joint})\right)\right)\right) & (5)\\ \boldsymbol{\Delta}\mathbf{S}_{exp} &= \mathrm{FC}^2_{exp}\left(\xi\left(\mathrm{FC}^1_{exp}(z_{exp})\right)\right) & (6)\\ &= \mathrm{FC}^2_{exp}\left(\xi\left(\mathrm{FC}^1_{exp}\left(\mathrm{FC}_{exp}(z_{joint})\right)\right)\right) & (7) \end{aligned}$$

with $\xi(x) = \max(0, x)$ the element-wise ReLU non-linearity.
Both decoders are symmetric, with $\mathrm{FC}^1_{(\cdot)} : \mathbb{R}^{512} \to \mathbb{R}^{1024}$ and $\mathrm{FC}^2_{(\cdot)} : \mathbb{R}^{1024} \to \mathbb{R}^3$.

Table 1 Summary of training data—reproduced from Liu et al. (2019)

Database	#Subj.	#Neu.	#Sample	#Exp.	#Sample
BU3DFE (Yin et al. 2006)	100	100	1000	2400	2400
BU4DFE (Yin et al. 2008)	101	> 101	1010	> 606	2424
Bosphorus (Savran et al. 2008)	105	299	1495	2603	2603
FRGC (Phillips et al. 2005)	577	3308	6616	1642	1642
Texas-3D (Gupta et al. 2010)	116	813	1626	336	336
MICC (Bagdanov et al. 2011)	53	103	515	–	–
BJUT-3D (Baocai et al. 2009)	500	500	5000	–	–
Real Data	1552	5224	17,262	7587	9405
Synthetic Data	1500	1500	15,000	9000	9000

3.2 Training Data

The training data is formed from seven publicly available face datasets of subjects from a wide range of ethnic backgrounds, ages, and gender, as well as a set of synthetic 3D faces. Table 1 summarizes the exact composition of the training set.

Synthetic faces Liu et al. (2019) use the BFM 2009 morphable model to synthesize neutral faces of 1500 subjects, and the 3DDFA expression model Zhu et al. (2015) to further generate 6 random expressions for each synthetic subject.

Real scans Both neutral and expressive scans are kept, and the data is unlabeled. The data was processed by first converting the scans to textured meshes using simple processing steps, e.g. Delaunay triangulation of the depth images. Automatic keypoint localization was applied on rendered frontal views of the scans to detect facial landmarks. The 2D landmarks were back-projected on the raw textured mesh using the camera parameters. The cropped BFM template was annotated with matching landmarks, such that Procrustes analysis could be applied to find a similarity transformation to align the raw scan with the template.

Pre-processing In Liu et al. (2019), the authors applied cropping to remove points outside of the unit sphere originating at the tip of the nose of the subject. The authors also applied mesh subdivision to obtain denser ground-truth meshes, thereby facilitating the sub-sampling of 29495 vertices from scans with insufficient native resolution. Finally, the sampling of points from the scans for training was done at the pre-processing stage. Data augmentation was carried out by randomly sampling vertices from some scans several times and storing the different point clouds separately.

3.3 Losses and Training Procedure

Liu et al. (2019) sample $N_s = N = 29495$ vertices from the (subdivided) scans. This number being equal to the number of vertices in the template is a choice, and not a requirement.

Since the synthetic scans are, by nature, in correspondence with the BFM template, Liu et al. (2019) use the element-wise ℓ_1 norm to train with supervision. For real scans, self-supervised training is carried out to minimize the Chamfer distance between the output $\mathbf{S}$ of the decoder and the potentially subdivided ground-truth scan.

Additional losses are used for synthetic and real scans. Edge-length loss is applied to discourage poor triangulations for the reconstruction. For real scans, the edge-lengths in the output are regularized towards those of the template. For synthetic scans, the edge-length loss is applied as a function of the difference between the edge-length of the input and the output meshes. Normal consistency is used for vertex normals. Due to the presence of noise in the raw scans in the mouth region (points from the inside of the mouth, teeth, or tongue), Laplacian regularization is applied to penalize large changes in curvature in a pre-defined mouth region on the BFM template.

The autoencoder is trained in successive phases. First, only the identity decoder is trained on the synthetic data only, then on a combination of synthetic and real data. After 10 epochs, the identity decoder and the fully-connected layer of the identity branch of the encoder are frozen (i.e. backpropagation is disabled) and the expression decoder is trained on synthetic data alone, and then on a mixture of synthetic and real data. Finally, both decoders and encoder branches are trained simultaneously on both synthetic and real scans. We refer the reader to the original work for details.

3.4 Limitations

We now study the limitations of the approach.

3.4.1 Data Processing and Representation

Cropping Although cropping is a simple solution to remove unnecessary parts of the scans, we argue relying on it makes the method less robust. Cropping points outside of the unit sphere centered at the tip of the nose is affected by the quality of the landmark detection. Similarly, choosing the unit sphere centered at the origin of the ambient space will be affected

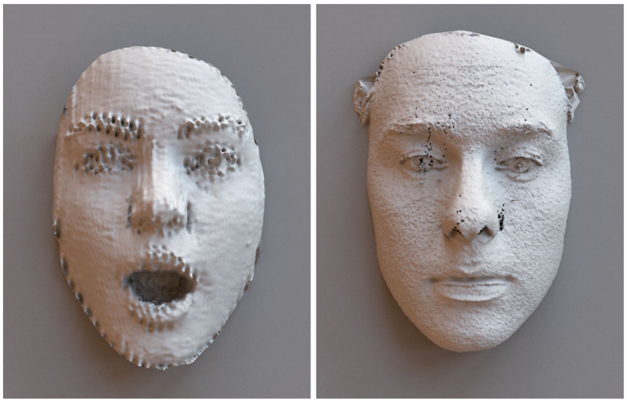


Fig. 2 Example sensor noise on the Bosphorus (left) and FRGC (right) datasets. Spikes highlighted on the FRGC scan

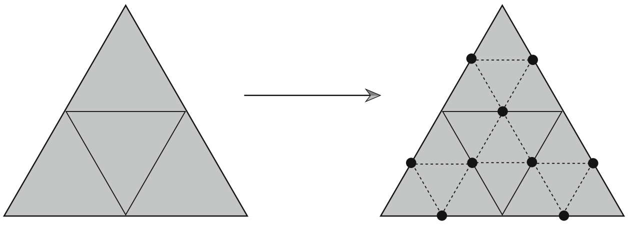


Fig. 3 Refinement step of the loop subdivision scheme. Adapted from (Pharr et al. 2016)

by the location of the scan in $\mathbb{R}^3$. In both cases, even though it is systematic, cropping is inconsistent: as the method is not adaptive, there is no guarantee that the noise (i.e. the points that do not contribute to a better face reconstruction and could even degrade the performance) will be discarded. In particular, for range scans such as those from the FRGC (Phillips et al. 2005), Bosphorus (Savran et al. 2008) and Texas 3D (Gupta et al. 2010) datasets, spikes an irregularities are commonly observed due to sensor noise, as shown in Fig. 2. Median filtering has traditionally been applied to the depth images before conversion to 3D surfaces as a means to alleviate this issue (Gupta et al. 2010), but incurs additional human intervention and might cause a loss of details. Cropping would not remove spikes, nor would it discard other irrelevant points if contained within the unit sphere. At the same time, cropping might discard points that would have contributed to the face region.

Subdivision scheme and vertex subsampling In Liu et al. (2019), mesh subdivision was used to improve the accuracy of the dense correspondences (i.e. provide more ground truth points for the Chamfer loss), and to enable consistent sampling of 29495 vertices for the input point cloud, even from low-resolution face scans that might not have enough remaining vertices in the facial region after cropping (e.g. most scans from the BU-3DFE database (Yin et al. 2006)). The authors then sampled 29495 vertices at random from the (subdivided) mesh to obtain a point cloud.

Subdivision schemes do not introduce additional details in the scan, but create a denser triangulation from existing triangles. The amount of memory required to store the same geometry is thus largely increased. Figure 3 illustrates the refinement step of the Loop scheme used by Liu et al. (2019). Assuming we started with one triangle and applied the scheme twice, the figure on the left in Fig. 3 shows the result after one subdivision step, and the figure on the right the result after two such steps. We can see that after one step, no vertices were introduced inside of the original triangle: all of the new vertices are located on its edges. After two steps, only 3 vertices have been placed inside the original triangle, yet the number of vertices has been multiplied by 5. In practice, two subdivision steps is the maximum that would be applied due to the rapid increase in memory required to store the subdivided meshes.

It is therefore apparent that a point cloud sampled uniformly at random from the vertices of the mesh cannot—in general—yield a uniform coverage of the surface, even after several mesh subdivision steps. Moreover, using the (subdivided) mesh as a ground truth in the Chamfer loss biases the reconstruction: closest points for vertices of the reconstructed mesh will either never be found inside the triangles of the scan, or in an unfavorable ratio when at least two subdivision steps have been applied.

Number of point clouds sampled per scan Liu et al. (2019) sampled one point cloud per expression scan, and *at most* ten point clouds per neutral scan, per subject. As this is done during pre-processing, all samples must be stored individually. No other data augmentation or transformation (e.g. jittering) was used. To avoid overfitting to a particular sampling of a given surface, we argue that as many different point clouds as possible should be presented to the model for each mesh.

3.4.2 Architectural Limitations and Conclusion

We review the limitations of the two main blocks of the algorithm of Liu et al. (2019), and conclude the section.

Decoder While MLP decoders are powerful and fully capable of representing details, they do not take advantage of the known template connectivity and geometry. In fact, careful tuning is required to obtain sound shapes: Liu et al. (2019) rely on a strong edge length prior, and use synthetic data extensively during training to condition both the encoders and decoders to respect the geometry of the template.

We observe significant artifacts for a large portion of the input scans, as shown in Fig. 4. Notably, we observe tearing-like artifacts and self-intersecting edges, as well as excessive roughness and ragged edges at the boundaries of the shape. In particular, heavy artifacting is present in the mouth region despite the use of the Laplacian loss. Such registrations cannot be exploited for downstream tasks (such as learning from

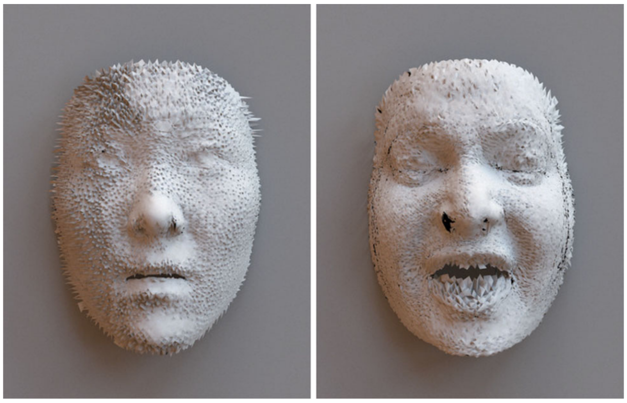
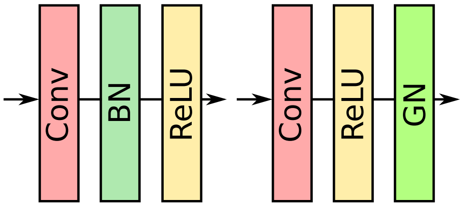


Fig. 4 Artifacts obtained with the architecture of Liu et al. (2019)



(a) PointNet **(b)** Our block

Fig. 5 Variants of the PointNet block: The vanilla PointNet block **a** consists of a 1 × 1 convolution followed by batch normalization and ReLU activation. We propose a variant **b** better suited to small batch sizes by replacing batch normalization with group normalization and normalizing the features post-activation

or statistical analysis on the registered scans) without heavy post-processing to correct the artifacts and improve surface fairness.

Encoder A vanilla PointNet (Qi et al. 2017a) layer consists of a 1 × 1 convolution, followed by batch normalization and a ReLU activation, as shown in Fig. 5a. Choosing $N_s = N$ facilitates mixed batching of synthetic and real scans, but according to Liu et al. (2019), the optimal batch size for the model was found experimentally to be 1. As batch normalization is known to result in degraded performance for small batch sizes (Wu and He 2020), we therefore investigate possible improvements.

*Number of parameters*While the PointNet encoder used in Liu et al. (2019) enables a high degree of weight sharing, the fully-connected decoders use dense fully-connected layers. This design choice results in a high number of parameters (183.6M), which, combined with the limited data augmentation and absence of regularization, promotes overfitting.

Conclusion The reliance on subdivision and cropping, the high number of trainable parameters, as well as the training methodology utilised, make the method of Liu et al. (2019) only suitable for in-sample registration, and thus the fast inference time does not fully offset the offline training time. The presence of significant noise and artifacts on registrations of scans from the training set further limits the applicability of the model on its own.

4 Description of the Method

We now introduce Shape My Face, our registration and morphable model pipeline. Our approach is based on the idea that registration can be cast as a translation problem, where one seeks to faithfully translate a latent geometric information (the surface) from an arbitrary input modality to a controlled template mesh. It is therefore natural to adopt an autoencoder architecture, with the advantages exposed in Sect. 3. We also wish to ensure our model is compact and performs reliably and satisfyingly on unseen data. The emphasis is, therefore, on robustness and applicability to real-world data, potentially on the edge.

4.1 Preliminaries and Stochastic Training

We choose the mean face of the LSFM model to be our template. We manually cropped the same facial region as the template of Liu et al. (2019) from a full-face combined LSFM and FaceWarehouse morphable model, and ensured a 1-to-1 correspondence between vertices. We choose LSFM since it is more representative of the mean human face than the BFM 2009 mean, and to facilitate the prototyping of a mouth model, as explained in Sect. 4.4.

We adopt a formulation in terms of blendshapes and define the output of our network to be

$$\mathbf{S} = \boldsymbol{\mu} + \boldsymbol{\Delta}\mathbf{S}_{id} + \boldsymbol{\Delta}\mathbf{S}_{exp} \tag{8}$$

where $\boldsymbol{\mu}$ is the template mean face shown in Fig. 6a, and $\boldsymbol{\Delta}_{id}$ and $\boldsymbol{\Delta}_{exp}$ are identity and expression deformation fields, respectively, defined on the vertices of $\boldsymbol{\mu}$. We motivate this choice to encourage better disentanglement by modeling both identity and expression as additive deformations of a plausible mean human face.

We follow an encoder-decoder architecture using a point cloud encoder and two symmetric non-linear decoders for the identity and expression blendshapes. As we will develop further, we propose a novel approach to avoid mouth artifacts by blending the non-linear blendshapes smoothly with linear blendshapes of the mouth region (defined based on the geodesic radius from the inside of the mouth). The flowchart of the method is presented in Fig. 7.

Input shape representations At inference time, our method only requires that we may randomly sample points on the surface of the scan. At training time, we optionally use the normal vectors at the sampled points (see Sect. 4.5). Therefore, any input modality that satisfies these requirements is suitable for training and inference.

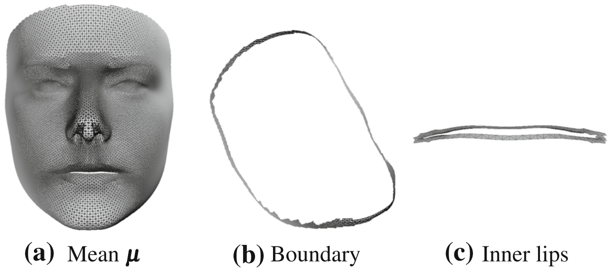

(a) Mean $\boldsymbol{\mu}$ **(b)** Boundary **(c)** Inner lips

Fig. 6 Parts of the face model: We decode shapes by predicting new vertex positions for the mean face of the LSFM model Booth et al. (2016, 2018a) (**a**). To avoid ragged boundaries, we encourage a small crop of the boundary (**b**) of the reconstructions to be close in position and curvature to that of the LSFM mean face. We propose a parameter-free approach for achieving high quality mouth reconstructions by reconstructing a crop of the mouth region on a small mouth-specific PCA model, and blending the reconstruction with the shapes predicted by the decoders using a smooth blending mask derived from the geodesic distance of the vertices in the template to a small crop of the lips (**c**)

In this work, we deal with training datasets of raw scans represented as meshes rigidly aligned (with scaling) with the template. Contrary to Liu et al. (2019), we do not apply any further processing on the 3D scans after rigid alignment. In particular, no surface subdivision and no offline sampling for data augmentation are done. We will also demonstrate inference on raw point clouds directly (Sect. 6.4). amically sample $N_s = 2^{16} = 65{,}536$ points uniformly at random on the surface of the input mesh using a triangle weighting scheme. Furthermore, we use the sampled point cloud as ground truth in the Chamfer loss. This ensures the vertices of the registration can be matched to points anywhere on the input surface, including inside triangles where the true projection of the vertices of the registration are more likely to lie.

We denote the triangulated raw input scan by the tuple $(\mathbf{S}_{in}, \mathbf{T}_{in})$, where $\mathbf{S}_{in}$ is the set of vertices of the mesh, and $\mathbf{T}_{in}$ the triangles. We write $\mathbf{P}_{in}$ the point cloud dynamically sampled on the surface of $(\mathbf{S}_{in}, \mathbf{T}_{in})$, and $\mathbf{N}_{in}$ the associated sampled point normals.

We use both synthetic and real scans in training. The training procedure is detailed in Sect. 4.6.

4.2 Encoder and Attention

In PointNet (Qi et al. 2017a), the authors introduce one of the first CNN architectures for point clouds. A PointNet layer consists of a 1 × 1 convolution followed by batch normalization and a ReLU activation, as shown in Fig. 5a. PointNet showed high performance on classification and segmentation tasks using moderately dense point clouds as input (2048 points for the ModelNet40 meshes). In this work, we sample $2^{16} = 65{,}536$ points from the input scans, which limits the batch sizes that can be accommodated with a single GPU

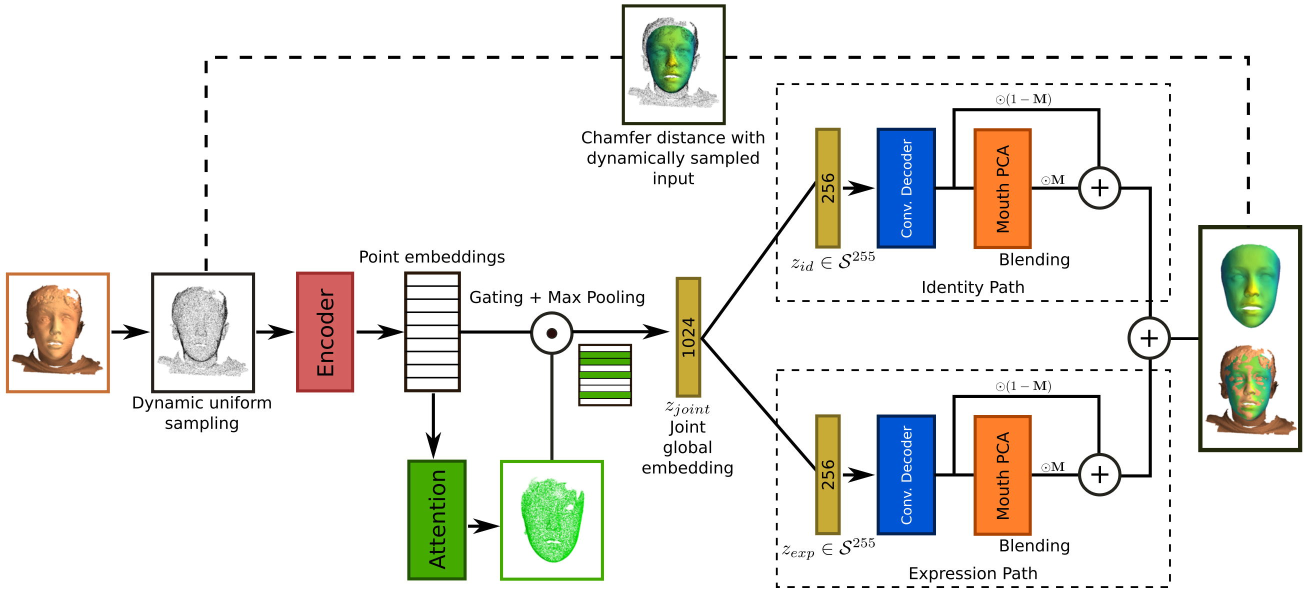


Fig. 7 Flow-chart representation of our approach: We sample 2^{16} points uniformly at random on the surface of the scan to register. A modified PointNet encoder computes features and an attention score for each point, from which a global embedding $\mathbf{z}_{joint}$ is obtained. We produce two hyperpsherical embeddings $\mathbf{z}_{id}$ and $\mathbf{z}_{exp}$ from $\mathbf{z}_{joint}$, and apply *mesh inception* decoders to output corresponding identity and expression blendshapes. To improve denoising, we smoothly blend the mouth region in a blendshape with its projection on a specialized PCA mouth model. During training (dotted lines), we measure the fit of the registration between the output of the network and the dynamically sampled input point cloud. This ensures vertices of the reconstruction can be matched to points anywhere on the surface of the scan, and not only to the vertices

implementation. As mentioned in Sect. 3.4.2, batch normalization is known to be ineffective for small batches (Wu and He 2020), as the sample estimators of the feature mean and standard deviation become noisy. We therefore propose modified PointNet layers with group normalization (Wu and He 2020), that we choose to apply after the ReLU non-linearity. Our modified PointNet layers are illustrated in Fig. 5b. We denote by $\mathrm{PN}(f_{in}, f_{out}, g)$ the block consisting of a 1×1 convolution with f_{in} input features and f_{out} output features, followed by one ReLU activation, and group normalization with group size g. The sequence of point convolutional layers in our encoder can thus be written $E(\cdot) = \mathrm{PN}(3, 64, 32) \rightarrow \mathrm{PN}(64, 64, 32) \rightarrow \mathrm{PN}(64, 64, 32) \rightarrow \mathrm{PN}(64, 128, 32) \rightarrow \mathrm{PN}(128, 1024, 32)$.

Visual attention To improve the robustness of our method to noise and variations in the physical extent of the scans, we introduce a novel visual attention mechanism implemented as a binary-classification PointNet sub-network applied to the features of the last PointNet layer and before the max-pooling operation. This can be seen as a form of region-proposal (He et al. 2017) or segmentation sub-network followed by a gating mechanism. We use our modified PointNet layers and obtain the following sequence of operations $\mathrm{PN}(1024, 128, 4) \rightarrow \mathrm{PN}(128, 32, 4) \rightarrow \mathrm{Conv}1 \times 1(32, 1)$. We use a smaller group size of 4 for group normalization to discourage excessive correlation in the features. The logits obtained as output of the attention sub-network are converted to a smooth mask by applying the sigmoid function and used as gating values to the max pooling operation—controlling which points are used to build the global latent representation $\mathbf{z}_{joint} \in \mathbb{R}^{1024}$ for the scan.

Hyperspherical embeddings Two dense layers predict separate identity and expression embeddings from $\mathbf{z}_{joint}$. We choose $\mathbf{z}_{id}, \mathbf{z}_{exp} \in \mathbb{R}^{256}$. Contrary to Liu et al. (2019), the mapping is non-linear: we normalize the identity and expression vectors, such that they lie on the hypersphere $\mathscr{S}^{255}$. Hyperspherical embeddings have been successful in image-based face recognition Wang et al. (2018); Deng et al. (2019) and shown to improve clusterability (Aytekin et al. 2018). Additionnally, we found the normalization to improve numerical stability during training.

The full encoder can be summarized as follows:

$$\tilde{\mathbf{Z}} = E(\mathbf{P}_{in}) \tag{9}$$

$$\mathbf{A} = \mathrm{Attention}(\tilde{\mathbf{Z}}) \tag{10}$$

$$\mathbf{z}_{joint} = \mathrm{MaxPool}(\sigma(\mathbf{A}) \odot \tilde{\mathbf{Z}}) \tag{11}$$

$$\mathbf{z}_{id} = \mathrm{Normalize}(\mathrm{FC}_{1024,256}(\mathbf{z}_{Joint})) \tag{12}$$

$$\mathbf{z}_{exp} = \mathrm{Normalize}(\mathrm{FC}_{1024,256}(\mathbf{z}_{Joint})) \tag{13}$$

where $\odot$ denotes the element-wise (Hadamard) product and $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function applied element-wise.

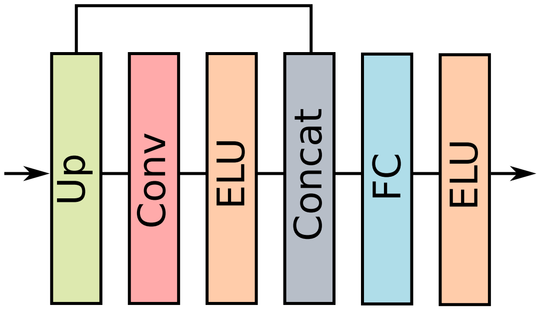


Fig. 8 One Mesh Inception block: Our mesh convolution block offers two paths for the information to flow from one resolution to the next. We concatenate the activated feature map of the current convolution layer with the upsampled feature map of the previous layer. The features are combined in a learnable way by a fully connected layer followed by another ELU activation

4.3 Mesh Convolution Decoders

As developed in Sect. 3.4.2, the fully-connected decoders used in Liu et al. (2019) suffer from two main challenges. First, they employ a high number of parameters, which promotes overfitting. Second, they do not leverage the known template geometry, and therefore require heavy tuning and regularization to produce sound shapes without abrupt changes in curvature and triangle geometry.

We propose non-linear decoders based on mesh convolutions. Our method is applicable to any intrinsic convolution operator on meshes. In this particular implementation, we use the SpiralNet++ operator. Denoting $\mathbf{x}_i^{(k)}$ the features of vertex i at layer k, we have:

$$\mathbf{x}_i^{(k)} = \gamma^{(k)} \left(||_{j \in S(i,M)} \, \mathbf{x}_j^{(k-1)} \right) \tag{14}$$

with $\gamma^{(k)}$ an MLP, $||$ the concatenation, and $S(i, M)$ the spiral sequence of neighbors of i of length (i.e. kernel size) M.

We observed training was difficult with the vanilla operators. As some operators such as SpiralNet++ and ChebNet already have a form of residual connections built-in (the independent weights given to the center vertex of the neighborhood), vanilla residuals or the recently-proposed affine skip connections (Gong et al. 2020) would be redundant. We instead propose a block reminiscent of the inception block in images (Szegedy et al. 2015) that can benefit any graph convolution operator. We concatenate the output of the previous upsampled feature map with the output of the convolution after an ELU non-linearity (Clevert et al. 2016). The concatenated feature maps are combined and transformed to the desired output dimension using an FC layer followed by another ELU non-linearity, as illustrated in Fig. 8.

We found this technique to drastically improve convergence and details in the reconstructed shapes. The technique is comparable to GraphSAGE (Hamilton et al. 2017), using graph convolutions followed by ELU as the AGGREGATE$_k$ function in (Hamilton et al. 2017, Algorithm 1), and ELU non-linearities. We refer to our block as *Mesh Inception*.

For upsampling, we follow the approach of Ranjan et al. (2018). We decimate the template four times using the Qslim method (Garland and Heckbert 1997) and build sparse upsampling matrices using barycentric coordinates. We set the kernel sizes of our convolution layers to 32, 16, 8, and 4, starting from the coarsest decimation of the template.

4.4 Mouth Model and Blending

Though the raw scans are rigidly aligned with the template on 5 facial landmarks that include the two corners of the mouth (Liu et al. 2019), the mouth expressions introduce a high level of variability in the position of the lips. Additionally, numerous expressive scans include points captured from the tongue, the teeth, or the inside of the mouth. This noise and variability in the dataset makes finding good correspondences for the mouth region difficult and leads to severe artifacting in the form of vertices from the lips being pulled towards the center of the mouth. In Liu et al. (2019), the authors advocate for the use of Laplacian regularization to prevent extreme deformations by penalizing the average mean curvature over a pre-defined mouth region, controlled by a weight λ_{Lap}. While this shows some success, we experimentally observed that, for small to moderate values of λ_{Lap}, artifacts remained. As shown in Fig. 9, while artifacts were reduced for large values of λ_{Lap}, so was the range of expressions.

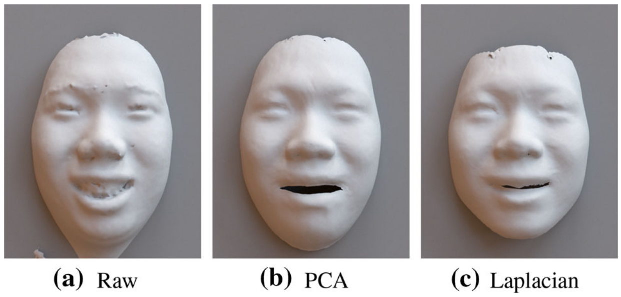
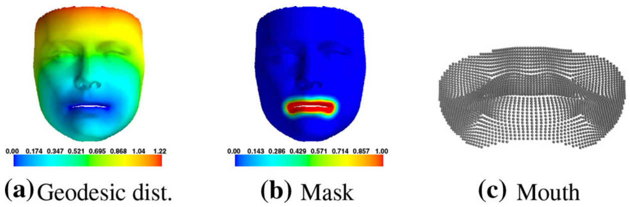

(a) Raw (b) PCA (c) Laplacian

Fig. 9 Laplacian loss and statistical mouth model: Laplacian loss (c) limits the expressivity of the scans but does not eliminate the artifacts completely (sample from the BU-3DFE dataset)



(a) Geodesic dist. (b) Mask (c) Mouth

Fig. 10 Mouth region and blending: From the small crop of the lips of Fig. 6c, we compute the geodesic distance of all vertices of the template to the vertices of the crop S_{inner} (**a**). We define the mouth region as the vertices within a chosen geodesic radius of S_{inner} (**c**). We define the blending mask as a function of the geodesic distance, shown as a heatmap in (**b**)

In this work, we introduce a new approach based on blending a specialized linear morphable model with the non-linear face model. We first isolate a small set of vertices, S_{inner}, from the innermost part of the lips of the cropped LSFM mean face, as shown in Fig. 6c. We then compute the geodesic distance from S_{inner} to all vertices of the template using the heat method with intrinsic Delaunay triangulation (Crane et al. 2017), which is visualised in Fig. 10a. We redefine the mouth region to be the set of vertices S_{mouth} within a given geodesic radius d from S_{inner}. By visual inspection, we choose $d = 0.15$. The resulting mouth region is shown as a point cloud in Fig. 10c.

To obtain a linear morphable model of this mouth region, we cropped the PCA components of the full face LSFM and FaceWarehouse model whose mean we used to obtain our face template. We keep only a subset, $\mathbf{W}_{id}$, of 30 identity components (from LSFM) and a subset, $\mathbf{W}_{exp}$, of 20 expression components (from FaceWarehouse). While it is well known that computing PCA on the cropped region of the raw data leads to more compact bases (Blanz and Vetter 1999; Tena et al. 2011), re-using the LSFM and FaceWarehouse bases enabled efficient prototyping. There is a trade-off between representation power and clean noise-free reconstructions: the model needs to be powerful enough to represent a wide range of expressions but restrictive enough that it does not represent the unnatural artifacts.

We project the mouth region of the blendshapes on the PCA mouth model during training and blend them smoothly with their respective source blendshapes, i.e., we project the mouth region of $\mathbf{S}_{id}$ on $\mathbf{W}_{id}$ and the mouth region of $\mathbf{S}_{exp}$ on $\mathbf{W}_{exp}$. Blending should be seamless, but—equally importantly—should also remove artifacts. We propose to define a blending mask intrinsically as a Gaussian kernel of the geodesic distance from S_{inner}:

$$b(r, c, \tau) = \begin{cases} \exp^{(-(r-c)^2/\tau^2)}, & \text{if } r \geq c \\ 1, & \text{otherwise.} \end{cases} \tag{15}$$

Where c and τ control the geodesic radius for which the PCA model is given a weight of 1, and the rate of decay, respectively. Compared to exponential decay, the squared ratio $((r-c)/\tau)^2$ allows us to favor more strongly the PCA model when $r - c \leq \tau$ and decay faster for $r - c > \tau$. Enforcing weights of 1 within a certain radius helps ensure the artifacts are entirely removed.

The mouth region of the blendshape $\mathbf{S}_{(.)}$ is redefined as:

$$\mathbf{S}_{(.),mouth} = \mathbf{M} \odot \left(\mathbf{P}_{(.)}\mathbf{Y}_{(.),mouth}\right) + (\mathbf{1} - \mathbf{M}) \odot \mathbf{Y}_{(.),mouth} \tag{16}$$

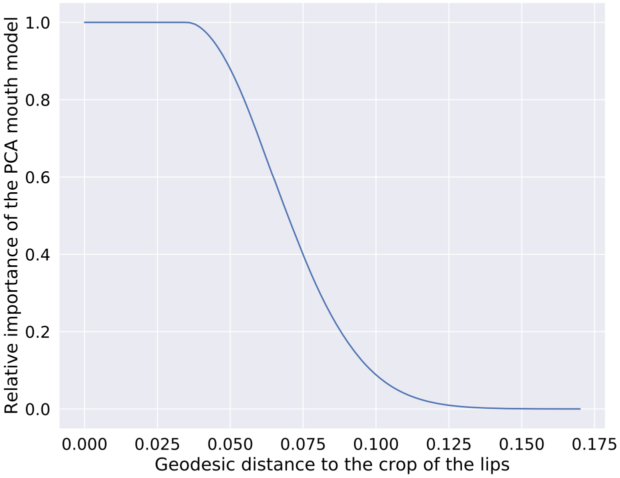


Fig. 11 Blending function: Plot of $b(r, c, \tau)$ for the values of c and τ used in this paper. We enforce a weight of 1 on the PCA model for the vertices within geodesic distance c of S_{inner}. We choose the rate of decay τ to enforce a weight close to 0 on the PCA model at the edges of the mouth region

With $\mathbf{M}$ the blending mask, $\mathbf{Y}_{(.),mouth}$ the mouth region in the output of the mesh convolutions, and $\mathbf{P}_{(.)}$ the projection matrix on the matching PCA basis.

We choose c experimentally. As c varies, we adapt τ to ensure the contribution of the PCA model to the reconstruction of the mouth region is low at the edges of the crop, and avoid seams. For a desired weight $\epsilon << 1$ at distance r and given c, we compute

$$\tau(r, c, \epsilon) = \frac{r - c}{\sqrt{-\log(\epsilon)}}. \tag{17}$$

In practice, we choose $c = 3.5e-2$ and $\epsilon = 5e-4$. We plot the resulting $b(\cdot, c, \tau)$ in Fig. 11.

In this work, we fixed c and τ for all shapes, on the assumption that the geodesic distance from the inner lips does not vary excessively in the dataset. However, it is perfectly reasonable to consider both parameters to be trainable, or to predict them from the latent vectors $\mathbf{z}_{joint}$, $\mathbf{z}_{id}$ or $\mathbf{z}_{exp}$ to obtain shape or blendshape-specific blending masks.

4.5 Losses

For synthetic scans, we define

$$L_{vertex}(\mathbf{S}, \mathbf{S}_{in}) = ||\mathbf{S}_{in} - \mathbf{S}||_1. \tag{18}$$

For real scans, we use the Chamfer distance

$$L_{vertex}(\mathbf{S}, \mathbf{P}_{in}) = \sum_{p \in \mathbf{S}} \min_{q \in \mathbf{P}_{in}} ||p - q||_2^2 + \sum_{q \in \mathbf{P}_{in}} \min_{p \in \mathbf{S}} ||p - q||_2^2. \tag{19}$$

As in Liu et al. (2019), we discard q from the error if

$$\min_{q \in \mathbf{P}_{in}} ||p - q||_2^2 > \sigma \quad \text{or} \quad \min_{p \in \mathbf{S}} ||p - q||_2^2 > \sigma. \tag{20}$$

We set $\sigma = 5e-4$.

For synthetic scans, we let $\mathbf{n}(p)$ be the normal vector at vertex $p \in \mathbf{S}$, and $\mathbf{n}_{in}(p)$ be the normal in the synthetic scan, and define the normal loss as:

$$L_{normal} = \frac{1}{N} \sum_{p \in \mathbf{S}} (1 - < \mathbf{n}(p), \mathbf{n}_{in}(p) >). \tag{21}$$

For real scans, we use

$$L_{normal} = \frac{1}{N} \sum_{p \in \mathbf{S}} (1 - < \mathbf{n}(p), \mathbf{N}_{in}(q) >), \tag{22}$$

where q is the closest point in $\mathbf{P}_{in}$ found by Eq. 19. In both cases, we set a weight of $\lambda_{norm} = 1e-4$.

Mesh convolutions are aware of the template connectivity and geometry, and do not require as much regularization as MLPs, we therefore use a weight of $\lambda_{edge} = 5e-5$ for the edge-loss, whose formulation is identical to Liu et al. (2019).

To regularize the attention mechanism during the initial supervised training steps, we assume all points sampled from the synthetic faces are equally fully important and none should be removed. We encourage the attention mask for the points sampled from synthetic scans to be 1 everywhere, using the binary cross entropy loss with a weight $\lambda_{att} = 1e-4$.

Finally, we enforce both an edge loss and ℓ_1 loss regularization between the reconstruction and the template in a small crop of the boundary, shown in Fig. 6b, to eliminate tearing artifacts. We let $\lambda_{bnd} = 1e-3$.

4.6 Training, Models, and Implementation Details

*Training data*As previously exposed, we use the same raw aligned data as the baseline model of Liu et al. (2019), but do not apply any further pre-processing, including data augmentation. To keep the ratio of identity and expression scans identical, we simply sample from the same scan as many times as required in a given training epoch.

In addition to the seven datasets of Table 1, we further add two large-scale databases of 3D human facial scans. The MeIn3D (Booth et al. 2017, 2018a; Bouritsas et al. 2019) database contains 9647 neutral face scans of people of diverse age and ethic background. We also select 17,750 scans from the 4DFAB (Cheng et al. 2018) database. 4DFAB contains neutral and expressive scans of 180 subjects captured in 4 sessions spanning a period of 5 years. Each session comprises up to 7 tasks, consisting of either utterances, voluntary, or spontaneous expressions.

Table 2 A very large scale morphable model: Summary of the additional databases used to train SMF+

Database	#Subj.	#Neu.	#Sample	#Exp.	#Sample
MeIn3D (Booth et al. 2017)	9647	9647	9647	0	0
4DFAB (Cheng et al. 2018)	180	6449	6449	11,301	11,301
Real Data (additional)	9727	16,096	16,096	11,301	11,301
Real Data (total)	11,379	21,320	33,358	18,888	20,706
Synthetic Data (total)	1500	1500	15,000	9000	9000

For a given subject in the 4DFAB database, we select the first frame of all sequences in the first two tasks as neutral scans. We select the middle frame of every sequence of the first two tasks as expressive scans for the six basic expressions (happy, sad, surprised, angry, disgust, and fear) and utterances. For tasks 3, 4, and 5, we select the frames at $1/3$ and $2/3$ of the sequence. For task 6, we select the frames at $1/3$ and $2/3$ of the sequence for the first two sessions, and the middle frame otherwise. We pick the middle frame for all other sequences.

In this work, we evaluate two models. We call SMF our model trained on the same dataset as the baseline. Our model trained with the addition of the MeIn3D and 4DFAB datasets is denoted by SMF+. The breakdown of the dataset for SMF+ is presented in Table 2.

Training procedure The BFM 2009 model was trained on a sample size of 200 subjects, and offers a limited representation of the diversity of human facial anatomy. We found the synthetic data to hinder the performance of the model, and to limit the realistic nature of the reconstructions. Mesh convolution operators learn to represent signals on the desired template and can readily exploit its connectivity and learn local geometric properties, we therefore drastically reduce the reliance on synthetic data to only the very first stages of training to condition the attention mechanism.

We first train the encoder and the identity decoder on synthetic data only for 5 epochs; and then on real neutral scans only for a further 10 epochs. We repeat this procedure for the expression decoder by freezing the identity decoder and the identity branch of the encoder, using only expressive scans. We then train both decoders jointly and the encoder for 10 epochs on the entire set of real scans. Finally, we change the batch size to 1 and refine the complete model for 15 epochs on the entire set of real scans.

We set the initial batch size to 2 and 8 for SMF and SMF+, respectively. We train the models with the Adam optimizer (Kingma and Ba 2014), with a learning rate of $1e-4$, and automatically decay the learning rate by a factor of 0.5 every 5 epochs. No additional regularization is used.

Software implementation and hardware Our model is implemented with Pytorch. We use the CGAL library for the computation of the geodesic distance using the heat method (Crane et al. 2020), implemented in C++ as a Pytorch extension. We render figures using the Mitsuba 2 renderer (Nimier-David et al. 2019).

All models were trained on a single Nvidia TITAN RTX, in a desktop workstation with an AMD Threadripper 2950X CPU and 128GB of DDR4 2133MHz memory.

Side by side comparison We summarize the differences between SMF and the baseline in Table 3.

5 Experimental Evaluation: Registration

We first evaluate SMF and SMF+ on surface registration tasks. In addition to the original data from Liu et al. (2019), we test the generalisability of our method on a previously unseen dataset, 3DMD. 3DMD is a high resolution dataset containing in excess of 24,000 scans captured from more than 3000 individuals. The dataset contains subjects from a wide range of ethnicities and age groups, each expressing a variety of facial expressions including neutral, happy, sad, angry and surprised. As stated in Sect. 3, we obtained a pre-trained model from Liu et al. (2019) trained on the entire dataset, which we use as a baseline.

5.1 Landmark Localization

To reproduce the experiment of Liu et al. (2019), we first evaluate the methods on the BU-3DFE database. We train SMF on the whole training set, as well as on the training set without BU-3DFE. We also re-trained a model using the methodology described in Liu et al. (2019), excluding BU-3DFE from the training set. We include SMF+ (in sample) for comparison. We also report the performance of Non-Rigid ICP (NICP) initialized with landmarks and with additional stiffness weights to regularize deformations of the boundary, and use the values reported in Liu et al. (2019) for the algorithms of Bolkart and Wuhrer (2015), Salazar et al. (2014), and Gerig et al. (2018). For 3DMD, we report the performance of NICP initialized with landmarks, the pre-trained model of Liu et al. (2019), SMF, and SMF+. For the sake of completeness, we also report the results of initializing NICP with the registration provided by SMF and SMF+ in place of the LSFM mean face and without landmarks information or stiffness weights. Given manual annotations on the raw

Table 3 Summary: Side by side comparison of SMF and the baseline of Liu et al. (2019)

	Baseline	SMF
Encoder	Vanilla PointNet	Modified PointNet
$\mathbf{z}_{joint}$ space	$\mathbb{R}^{1024}$	$\mathbb{R}^{1024}$
$\mathbf{z}_{id}$ space	$\mathbb{R}^{512}$	$\mathscr{S}^{255} \subset \mathbb{R}^{256}$
$\mathbf{z}_{exp}$ space	$\mathbb{R}^{512}$	$\mathscr{S}^{255} \subset \mathbb{R}^{256}$
# input points	29,495	65,536
Template	BFM 2009	LSFM
Decoders	2-layer MLPs	Mesh inception
Preprocessing	Cropping, Subdivision, Data augmentation	None
Input	Pre-computed	Stochastic
Ground truth	Subdivided mesh	Stochastic
Losses	ℓ_1/ Chamfer, normal, edge, Laplacian	ℓ_1/ Chamfer, normal, edge, boundary, attention
Additional features	None	Visual attention
Trained model file size	701MB (`float32`)	179MB (`float32`)
# Trainable params.	183.6 millions (100%)	15.5 millions (8.8%)

scans, grouped by semantic label $(l_i^*)_{i=1}^k$ (e.g. left eye, or nose), we compute the semantic landmark error per landmark group as $\frac{1}{k}\sum_{i=1}^k ||\hat{l}_i - l_i^*||$, with $\hat{l}_i$ the corresponding landmark in the registration.

We report the mean and standard deviation of the error within each group. Table 4 summarizes the results for the BU-3DFE dataset, and Table 5 the results on 3DMD. We note that applying NICP did not significantly change the landmark error, which is likely due to the reconstructions output by SMF and SMF+ being already sufficiently close to the ground truth surface. There is, however, an advantage in using SMF to initialize NICP compared to landmarks: the typical runtime of the public implementation of NICP we used (from the publicly available LSFM code (Booth et al. 2018a)) with landmarks initialization was between 45 and 60 seconds per scan, while the initialization with SMF achieved equally detailed registrations in around 20 seconds per scan.

We note the high error values for the jaw landmarks on both datasets. The landmarks for the chin and jaw are at the boundary of the template. Since our method is trained on point clouds sampled from the raw scans with no manual cropping, the closest points for the boundary of the template are not at the edges of a tight crop of the face, and therefore the vertices of the boundary get pulled farther than where jaw landmarks are manually annotated. This results in large error values for these landmarks. Increasing the weight of our boundary loss regularization may help mitigate this phenomenon.

5.2 Surface Error

While a low landmark localization error suggests key facial points are faithfully placed in the registration, it does not paint the whole picture and does not indicate the general reconstruction fidelity. In particular, it is affected by the inevitable imprecision of manual labeling, and the error is measured on a small number of points.

To further assess the performance of the models, we measure the surface reconstruction error between the registrations and the ground-truth raw scans. We randomly select a sample of 5000 training scans and a sample of 5000 test scans (from the 3DMD dataset) and measure the distance of the vertices of the registrations to their closest point anywhere on the ground-truth surface (i.e., not the closest vertex). We summarize each scan by its mean surface error.

*Training and test set*We visualize the error distribution on the subsets of the training and test sets in Fig. 12. On both the training and test sets, the models exhibit typically low error, with a pronounced skew of the mean towards lower values (0.306 mm for SMF and 0.297 mm for SMF+). On the training set, the mean (per scan) error distributions of SMF and SMF+ appear very similar, with a slight advantage to SMF+. On the test set, however, SMF+ displays significantly lower values at the quartiles and a tighter distribution, suggesting the addition of the MeIn3D and 4DFAB datasets was effective in reducing the generalization gap and the variance of the model.

BU-3DFE and 3DMD To complete the evaluation on BU3D and 3DMD, we produce in Fig. 13 the cumulative distribution function (CDF) of the surface error for the entire BU3D dataset, and for the aforementioned sample of 5000 test scans, for the same models as in Sect. 5.1. To help visualize the counts of extreme values, we provide a rug plot for SMF evaluated out of sample. As evidenced by the plots, SMF and SMF+ performed very similarly, while SMF trained without BU-3D had slightly lower performance. The baseline model,

Table 4 Semantic landmarks error on BU-3DFE: Comparison of the mean and standard deviation for semantic landmark error (mm) for BU-3DFE using the *BU-3DFE* 83 facial landmark set

Region	NICP	GMCO	FAEIFC	GPMM	Baseline In	Baseline Out	SMF In	SMF Out	SMF+ In
L Eyebrow	4.59 ± 1.34	6.28 ± 3.30	4.69 ± 4.64	6.25 ± 2.58	7.98 ± 2.77	19.75 ± 4.09	7.20 ± 2.15	6.59 ± 1.99	7.23 ± 2.24
R Eyebrow	4.37 ± 1.32	6.75 ± 3.51	5.35 ± 4.69	4.57 ± 3.03	6.84 ± 2.51	20.86 ± 4.25	6.70 ± 2.23	6.48 ± 2.19	6.81 ± 2.31
L Eye	3.28 ± 0.98	3.25 ± 1.84	3.10 ± 3.43	2.00 ± 1.32	5.08 ± 1.65	11.75 ± 1.66	3.50 ± 1.10	3.40 ± 1.09	3.92 ± 1.22
R Eye	3.09 ± 0.98	3.81 ± 2.06	3.33 ± 3.53	2.88 ± 1.29	3.80 ± 1.35	10.27 ± 1.59	4.82 ± 1.44	5.09 ± 1.45	4.92 ± 1.46
Nose	3.74 ± 0.87	3.96 ± 2.22	3.94 ± 2.58	4.33 ± 1.24	4.90 ± 1.24	9.14 ± 1.56	4.62 ± 1.20	4.64 ± 1.17	4.48 ± 1.19
Mouth	4.82 ± 2.33	5.69 ± 4.45	3.66 ± 3.13	4.45 ± 2.02	5.32 ± 2.28	8.56 ± 2.40	6.25 ± 2.39	6.00 ± 2.36	6.33 ± 2.39
Chin	7.75 ± 2.79	7.22 ± 4.73	11.37 ± 5.85	7.47 ± 3.01	11.39 ± 4.78	25.69 ± 6.93	38.73 ± 9.18	38.08 ± 8.47	38.96 ± 9.88
L Face	8.14 ± 2.34	18.48 ± 8.52	12.52 ± 6.04	12.10 ± 4.06	15.63 ± 6.09	29.30 ± 7.35	14.08 ± 3.64	12.97 ± 3.50	14.31 ± 3.79
R Face	7.68 ± 1.91	17.36 ± 9.17	10.76 ± 5.34	13.17 ± 4.54	11.93 ± 4.02	27.38 ± 5.11	20.61 ± 5.19	19.59 ± 5.23	20.64 ± 5.34
Avg Face	4.12 ± 0.83	–	–	–	5.79 ± 1.34	13.40 ± 1.89	5.70 ± 1.11	5.52 ± 1.09	5.77 ± 1.12
Avg	4.91 ± 0.80	8.49 ± 4.29	8.09 ± 5.75	6.52 ± 3.86	7.37 ± 1.57	16.45 ± 2.19	9.07 ± 1.28	8.72 ± 1.24	9.15 ± 1.36

Landmark regions are as described in Salazar et al. (2014). *L* and *R* are shorthand for *Left* and *Right* respectively. *Avg Face* is the average for all inner face landmarks, and therefore excludes *Chin*, *L Face*, and *R Face*. Baseline is Liu et al. (2019), GMCO is Bolkart and Wuhrer (2015), FAEIFC is Salazar et al. (2014), and GPMM is Gerig et al. (2018)

Table 5 Semantic landmarks error on 3DMD: Comparison of the mean and standard deviation for semantic landmark error (mm) for 3DMD using the *ibug* 68 facial landmark set

Region	NICP	Liu et al. (2019)	SMF	SMF+	SMF + NICP	SMF+ + NICP
L Eyebrow	5.94 ± 2.16	6.23 ± 1.54	5.57 ± 1.66	5.87 ± 1.84	5.54 ± 1.70	5.84 ± 1.86
R Eyebrow	5.27 ± 2.05	6.14 ± 1.74	6.32 ± 2.20	6.70 ± 2.18	6.27 ± 2.21	6.68 ± 2.19
L Eye	4.29 ± 1.26	3.83 ± 1.17	4.27 ± 0.99	4.75 ± 0.99	4.25 ± 1.00	4.79 ± 0.99
R Eye	4.02 ± 1.37	3.79 ± 1.27	4.03 ± 1.18	4.08 ± 1.15	3.98 ± 1.19	4.01 ± 1.16
Nose	4.56 ± 0.96	4.94 ± 1.17	5.30 ± 0.85	5.42 ± 0.87	5.22 ± 0.83	5.32 ± 0.86
Mouth	3.96 ± 1.70	4.73 ± 1.65	6.36 ± 1.16	6.38 ± 1.15	6.31 ± 1.16	6.25 ± 1.14
Jaw	24.58 ± 4.69	35.76 ± 5.59	24.91 ± 5.67	25.25 ± 5.75	24.87 ± 5.73	25.22 ± 5.76
Avg Face	4.43 ± 0.95	4.84 ± 1.05	5.57 ± 0.74	5.73 ± 0.76	5.52 ± 0.73	5.65 ± 0.74
Avg	9.47 ± 1.59	12.57 ± 1.76	10.41 ± 1.75	10.61 ± 1.78	10.36 ± 1.76	10.55 ± 1.77

L and *R* are shorthand for *Left* and *Right* respectively. *Avg* is the average for all inner face landmarks

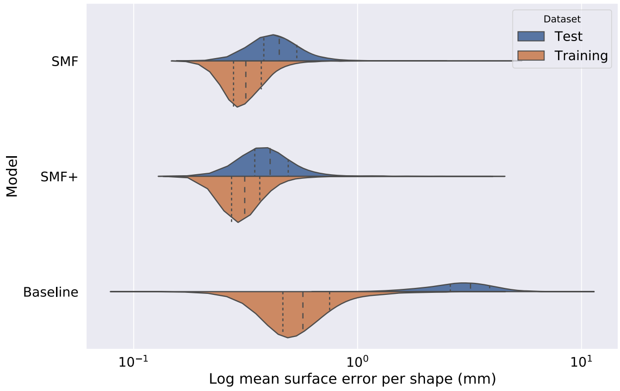


Fig. 12 Visualizing the error mean distribution on the training and test sets: violin plots of the mean (per scan) surface error on the training and test sets for SMF, SMF+, and the baseline, plotted on a log scale. A violin plot represents the range of the data along with a kernel density estimation of the distribution. We split the plots to help compare the error distribution on the two datasets. Vertical dotted lines represent the quartiles of the distribution

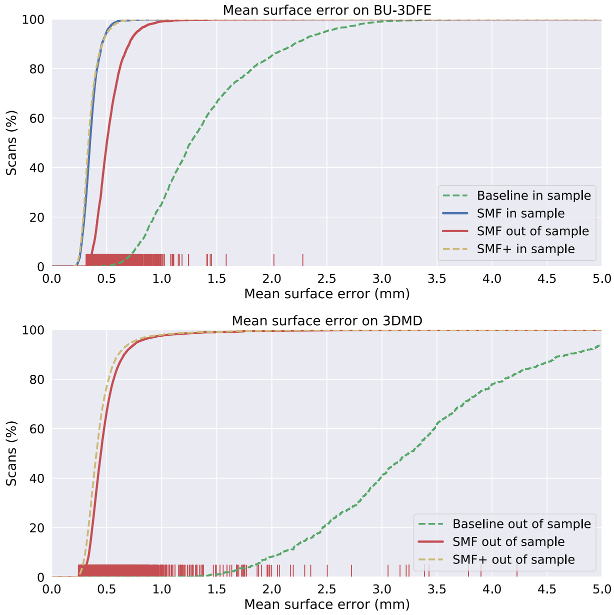


Fig. 13 Cumulative error: Cumulative distribution function for the mean (per scan) error on the training and test sets for the models evaluated for the semantic landmark accuracy experiment. Even though the semantic landmark error of the baseline was not atypical, the distribution of the surface error reveals that the registration accuracy is actually much lower than that of SMF and SMF+. The rug plot (red bars at the bottom) visualize the distribution of the samples in terms of mean error for SMF evaluated out of sample. On 5000 sample test scans, few outliers had high mean surface error. SMF+ performs comparably with SMF in sample but has distinctly lower generalization error

on the other hand, had significantly worse error distribution. Table 6 provides numerical values for the 25%, 50%, 75%, and 99% quantiles for the models we plotted. We omit the baseline evaluated out of sample on BU-3DFE due to the very high landmark localization error. On our separate test set, a similar development unfolds.

The difference between the error distributions of SMF+ and SMF is small but significant, with SMF+ outperforming the model trained on less data. Out of sample, the baseline model's performance is significantly degraded, with the bottom 25% of the surface error already reaching 2.60mm.

Training and test reconstructions We visualize sample reconstructions from SMF on the training and test sets. For each scan, we render the input point cloud sampled on the mesh, and the attention score predicted by SMF for every point as a heatmap, with bright green denoting attention scores close to 1, and black denoting attention scores close to 0. We also render the reconstruction produced by SMF, and the heatmap of the surface error as a texture on the registration. We render the reconstruction produced by the baseline for comparison. Figure 14 provides visualizations for 18 training samples arranged in two columns. Figure 15 shows the comparative performance of the baseline and our model for 12 test scans arranged in two columns. We show sample reconstructions from SMF+ on MeIn3D and 4DFAB in Fig. 16.

Table 6 Mean surface error quantiles: on BU-3DFE (left) and 3DMD (right) in mm

BU-3DFE	25%	50%	75%	99%	3DMD	25%	50%	75%	99%
SMF In	0.306	0.347	0.396	0.597	SMF Out	0.381	0.447	0.535	1.489
SMF+ In	0.297	0.333	0.387	0.617	SMF+ Out	0.347	0.407	0.490	1.326
SMF Out	0.434	0.501	0.594	0.983	–	–	–	–	–
Baseline In	0.995	1.261	1.683	2.989	Baseline Out	2.605	3.239	3.894	6.497

Visual inspection correlates strongly with the numerical evaluation. Our SMF model consistently produces registrations that are smooth and detailed, with very low surface error. The attention mechanism appears to successfully segment the face, eliminating gross corruption, and discarding points from the tongue and teeth for several scans. Our model faithfully represents both the identity and expression, even for extreme expressions on the test set.

In particular, factors such as age, ethnicity, and gender are accurately captured. Non-linear deformations of the nose, cheeks, and mouth are well preserved across a wide range of identities and expressions. Finally, despite the inclusion of points from the teeth and tongue in the raw scans, SMF produces artifact-free and expressive mouth reconstructions with seamless blending in the vast majority of cases.

5.3 Stability to Resampling

Given the stochastic nature of the method, we evaluate the stability of the reconstructions to resampling of the input scans. We then focus on evaluating the attention mechanism.

We select a subset of 1000 scans each of the training and test sets and produce 100 reconstructions with SMF, randomly sampling a new point cloud on the surface of the scan at each iteration. For each scan, we compute the mean reconstruction. For each point of the 100 reconstructions, we compute its Euclidean distance to the matching point in the mean reconstruction for that scan. We then take the median and max of these distances for every point in the scan and compute their median across the scan, denoted by “median median” and “median max”, as indications of the typical typical-case and typical worst-case variations. We collect both values for each of the 1000 training and 1000 test scans, and plot their histograms in Fig. 17. The results show our method is stable with respect to resampling, the median median variations, in particular, are concentrated below 0.1 mm with a typical maximum variation in position from the mean below 0.2 mm per vertex. Interestingly, we observe less spread on the test set than on the training set, but slightly higher typical maximum displacement per vertex, still below 0.2 mm per vertex. Figure 18 illustrates that the attention mechanism is also stable.

5.4 Ablation Study on the Decoder

We now study different variations of SMF by changing the decoder. Figure 19 presents the comparative performance of SMF, SMF+ and the ablations measured by average surface reconstruction error and ordered by test error. We also report the landmark localization errors of some of the variants in Table 7 for BU-3DFE and Table 8 for 3DMD.

5.4.1 Single Decoder

While two or more decoders can be used to promote separation between factors of variation in the data, and ties to the morphable model and generative model aspect of our work, our registration framework is equally applicable to single decoders (abbreviated s.d.). We keep the architecture of Sect. 4.3 and the mouth models of Sect. 4.4, and set the dimension of the latent space to 256 (“SMF s.d.”) and 512 (“SMF 512 s.d.”). As can be seen in Fig. 19, SMF and SMF 512 s.d. have similar average error and error variance, with SMF 512 s.d. slightly outperforming SMF, while SMF s.d. shows a slightly greater drop in performance. These results are expected: training a single decoder is no harder than training two and using a single mouth model with both identity and expression bases is also simpler, but the single latent space of dimension 256 in SMF s.d. further constraints the model compared to SMF. Sample reconstructions are presented in Fig. 20.

5.4.2 Fully-Connected Decoders

We investigate whether the improvements in the encoder and training methodology enable generalization with fully-connected decoders, and how the performance of such decoders compares to that of our mesh convolutional decoders. We follow the same architecture as Liu et al. (2019) for the decoders. The models compared are: SMF fc, obtained by substituting the mesh inception decoders with MLPs, keeping the dimension of the identity and expression latent spaces to 256 and all other hyperparameters identical and SMF 512 fc with latent spaces of dimension 512.

As reported in Fig. 19, SMF outperforms both variants in terms of surface error, while the fc variants performed better in terms of landmark error. Visualizing the reconstructions in

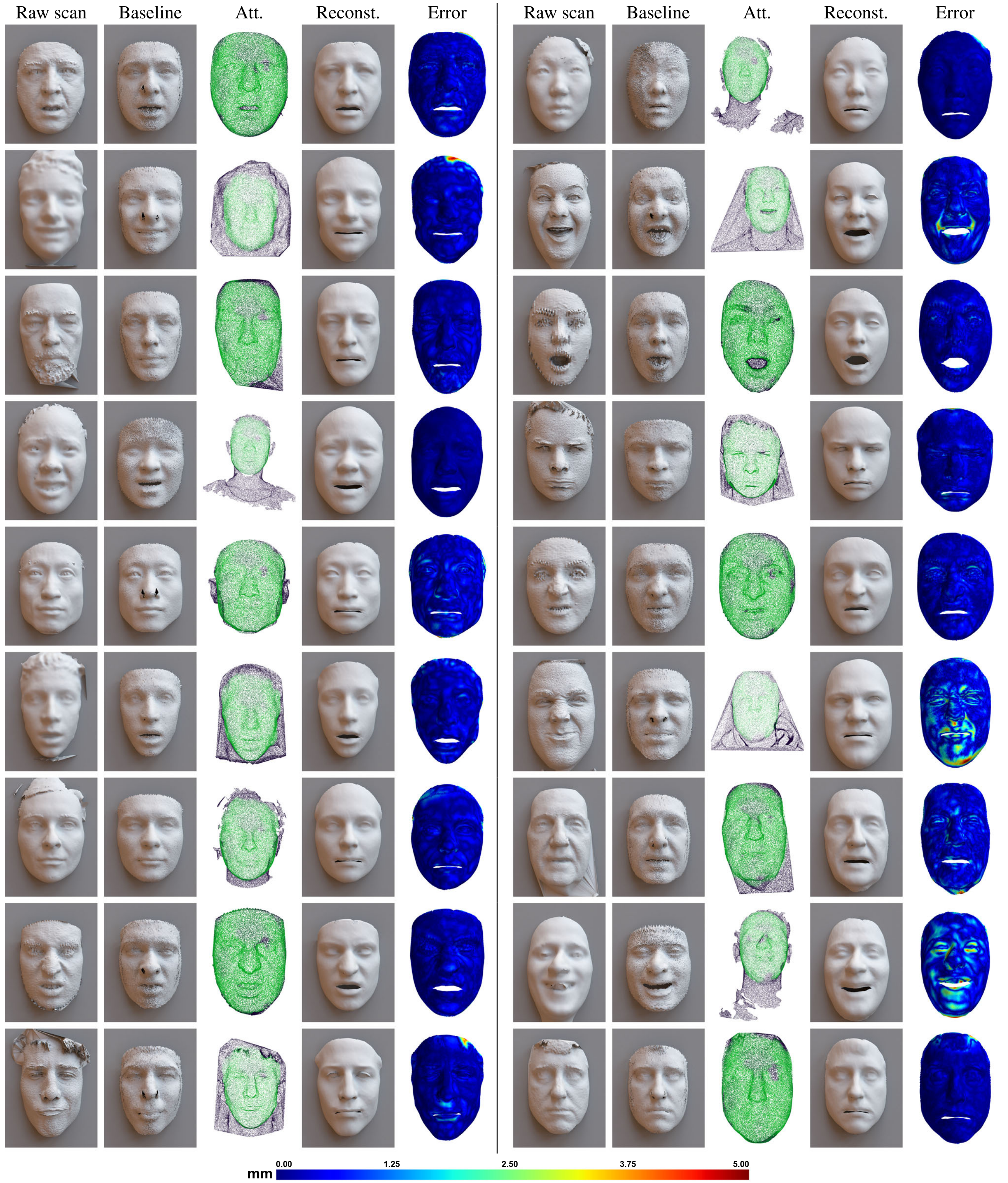


Fig. 14 Sample reconstructions on the training set for SMF: arranged in two columns. From left to right: raw scan, output of the baseline, point cloud sampled on the scan by SMF and predicted attention mask, output of SMF, and surface reconstruction error visualized as a texture on the output of SMF. We can see SMF markedly outperforms the baseline and provides accurate natural-looking reconstructions with uniformly low error in the facial region and accurate representation of both identity and expression

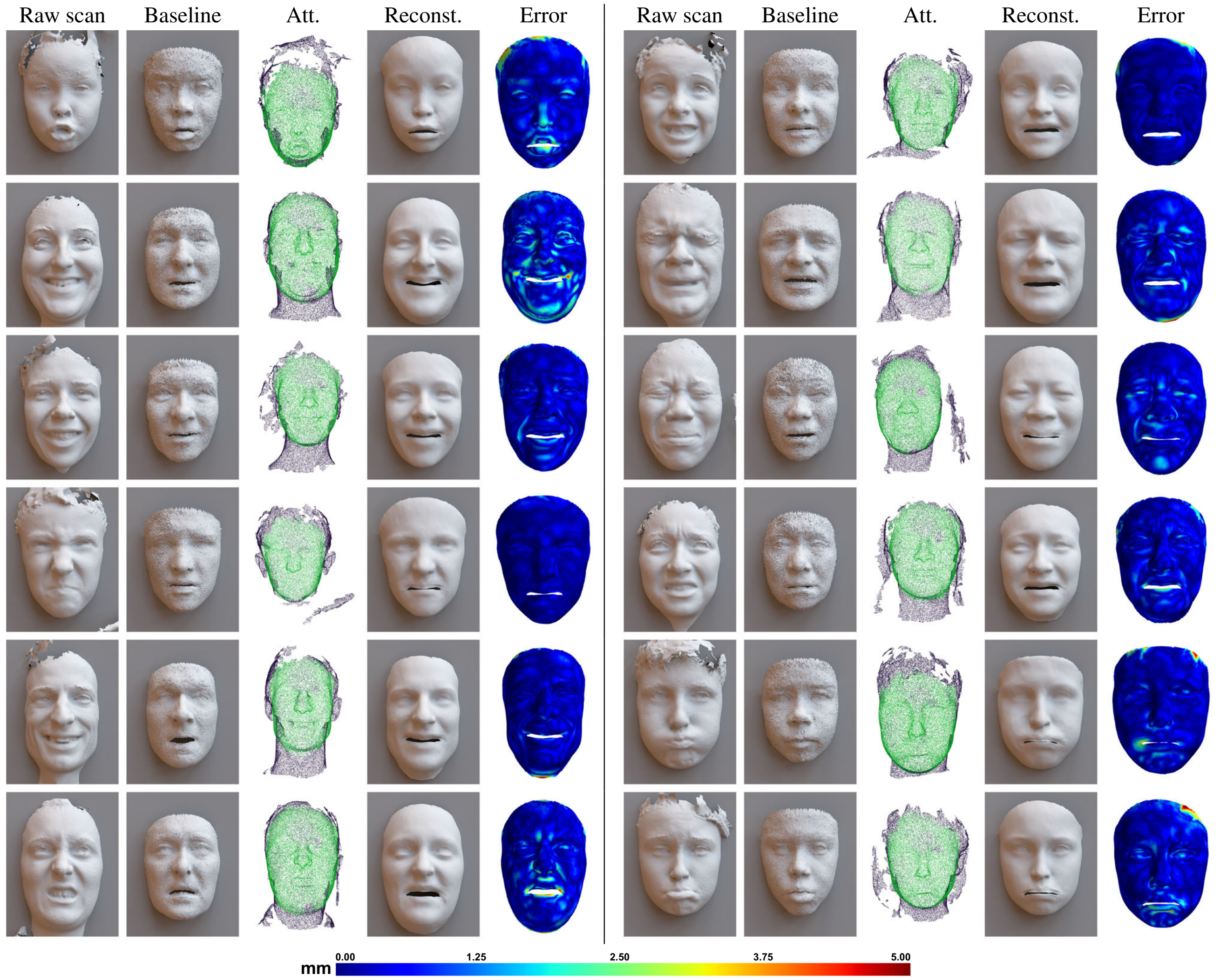
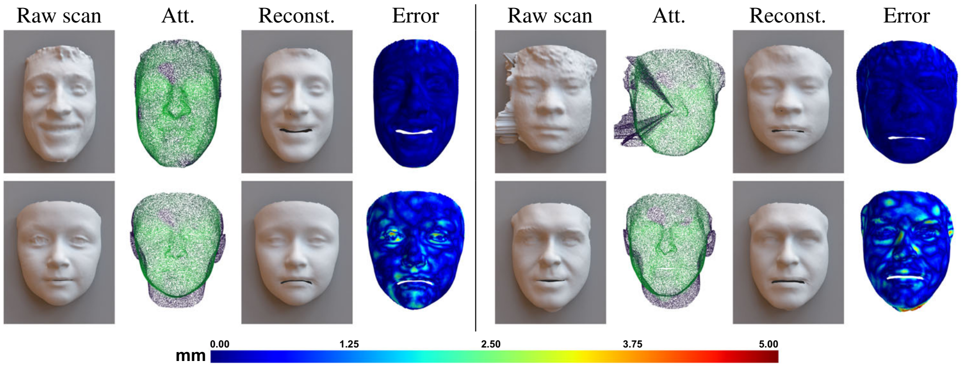


Fig. 15 Sample reconstructions on the test set for SMF: arranged in two columns. From left to right: raw scan, output of the baseline, point cloud sampled on the scan and predicted attention mask, output of SMF, and surface reconstruction error visualized as a texture on the output of SMF. The test reconstructions look comparable to the training reconstructions for SMF, with high quality registrations across gender, age and ethnicity, even for extreme facial expressions

Fig. 16 Sample reconstructions on additional training data for SMF+: arranged in two columns. From left to right: raw scan, point cloud sampled on the scan by SMF+ and predicted attention mask, output of SMF+, and surface reconstruction error visualized as a texture on the output of SMF+. Top row: 4DFAB, bottom row: MeIn3D



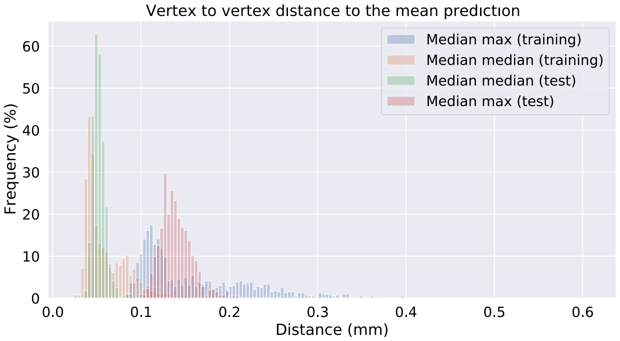


Fig. 17 Per vertex distance to the mean prediction: We sample 100 different point clouds for 1000 training and test scans and compute, for each vertex in each registration, its median Euclidean distance to the matching vertex in the average reconstruction. We present histograms of the max and median values (across vertices) per scan to show our method is stable to resampling the same input surface

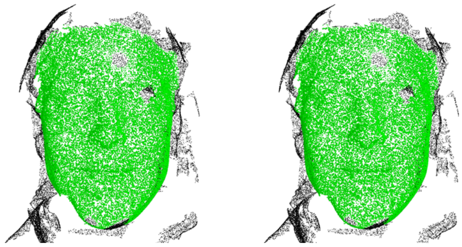


Fig. 18 Attention mask: Attention mask for two point clouds sampled from the same *test* shape (3DMD). It can be seen the attention mechanism excludes the points inside of the mouth and outside of the face area. The mask is also stable to resampling of the scan

Fig. 20, however, shows heavy noise. We therefore increased the edge-length regularization to $\lambda_{edge} = 2e - 4$ and re-evaluated the models (now SMF fc' and SMF 512 fc'). The models with increased edge-length regularization provided smoother reconstructions, but still suffered from artifacting and also performed worse in both surface error and landmark error. This ablation confirms that, in order to obtain reconstructions that are free from noise and large variations in curvature with fully-connected decoders, increased regularization is required, at some cost in accuracy. It is also apparent that some error metrics, such as landmark localization error, favor models that fit the positions of individual vertices at the expense of surface fairness.

It is worth noting that SMF with fully-connected decoders generalizes well to the test set, and that fully-connected decoders may in some cases provide finer details, albeit with additional noise. Our mesh inception decoders, however, achieve comparable performance, with no noise and with a fraction of the trainable parameters. We also note that the variance of the mean surface error is higher with fully-connected decoders, as indicated by the wider error bars in Fig. 19.

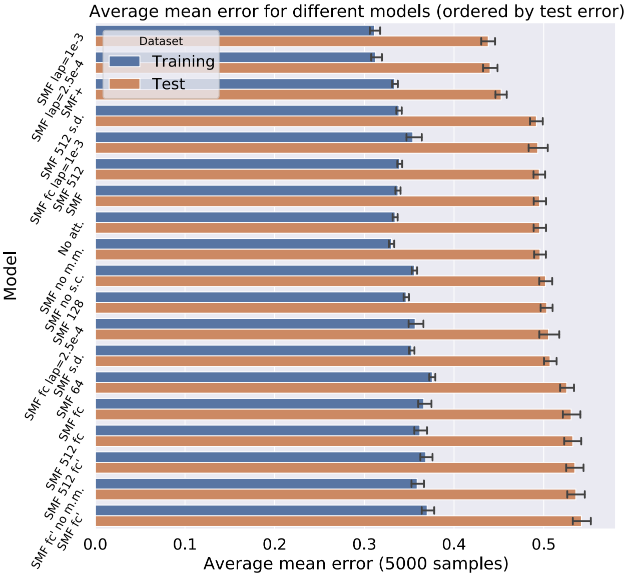


Fig. 19 Comparison of the average mean (per scan) surface fitting error for different choices of decoders, on 5000 random training scans and 5000 random test scans, ordered by average test error

5.4.3 Skip Connections

We now compare our mesh inception decoders with standard SpiralNet++ decoders, keeping all hyperparameters equal, and report the performance of "SMF no s.c.". The model without skip connections performed noticeably worse than SMF in terms of surface error, and landmark error on BU-3DFE, but was slightly better on the landmark localization task on 3DMD. Visual inspection in Fig. 20 reveals the presence of artifacts, especially around the mouth area.

5.4.4 Mouth Model

As stated in Sect. 4.4, the purpose of introducing a constrained PCA model for the mouth is to produce reconstructions that do not display unnatural deformations of the mouth in the presence of noisy points from the teeth or tongue, by finding a trade-off between model expressivity and robustness. Thus, it is expected that using the PCA model may come at a loss of precision.

We compare the models with mesh inception and fully-connected decoders in four scenarios: PCA mouth model (SMF, SMF fc'), no mouth model and no regularization (SMF no m.m., SMF fc' no m.m.), and Laplacian regularization with weight $\lambda_{lap} = 2.5e - 4$ and $\lambda_{lap} = 1e - 3$. We implemented Laplacian regularization using uniform weighting, as we found cotangent weights to be highly numerically unstable, leading to severe artifacting in the mouth region. We report the average surface reconstruction errors of the models in Figure 19 as well as their landmark localization errors

Table 7 Semantic landmarks error on BU-3DFE for the ablations compared: Comparison of the mean and standard deviation for semantic landmark error (mm) for BU-3DFE using the *BU-3DFE* 83 facial landmark set

Region	SMF	SMF no s.c.	SMF fc	SMF fc'	SMF fc' no m.m.	SMF fc lap=1e-3	SMF no m.m.	SMF lap=1e-3	SMF s.d.	SMF 512 s.d.
L Eyebrow	7.20 ± 2.15	8.56 ± 2.50	7.79 ± 2.14	6.89 ± 2.19	7.14 ± 2.28	8.27 ± 2.41	6.98 ± 2.03	7.93 ± 2.24	7.13 ± 2.18	6.57 ± 2.02
R Eyebrow	6.70 ± 2.23	8.63 ± 2.60	8.46 ± 2.51	7.00 ± 2.38	7.16 ± 2.35	6.85 ± 2.19	6.47 ± 2.13	6.26 ± 1.96	6.81 ± 2.32	6.42 ± 2.21
L Eye	3.50 ± 1.10	8.15 ± 1.56	4.97 ± 1.35	3.80 ± 1.24	3.73 ± 1.23	4.71 ± 1.37	3.63 ± 1.11	4.77 ± 1.25	3.46 ± 1.13	3.38 ± 1.10
R Eye	4.82 ± 1.44	8.11 ± 1.62	6.59 ± 1.69	5.06 ± 1.58	4.97 ± 1.59	5.83 ± 1.64	5.05 ± 1.43	4.36 ± 1.24	5.76 ± 1.57	4.04 ± 1.32
Nose	4.62 ± 1.20	5.93 ± 1.41	5.84 ± 1.16	4.38 ± 1.01	4.59 ± 1.01	5.13 ± 1.16	4.51 ± 1.18	4.51 ± 1.21	4.84 ± 1.18	4.73 ± 1.11
Mouth	6.25 ± 2.39	6.27 ± 2.13	5.62 ± 2.22	5.17 ± 2.37	5.84 ± 2.35	5.67 ± 2.20	9.54 ± 2.38	6.62 ± 2.29	5.96 ± 2.39	6.09 ± 2.42
Chin	38.73 ± 9.18	38.57 ± 8.67	33.38 ± 8.39	27.87 ± 8.18	28.27 ± 8.11	16.40 ± 6.84	37.24 ± 9.14	17.84 ± 6.61	31.74 ± 8.70	34.01 ± 8.82
L Face	14.08 ± 3.64	13.81 ± 3.41	13.94 ± 5.01	12.00 ± 4.45	11.73 ± 4.38	13.57 ± 5.23	13.96 ± 3.63	11.99 ± 4.25	12.83 ± 3.57	13.43 ± 3.66
R Face	20.61 ± 5.19	18.92 ± 4.75	14.23 ± 3.39	13.33 ± 3.86	13.44 ± 3.88	12.51 ± 3.65	19.80 ± 5.10	13.20 ± 3.86	17.97 ± 4.90	19.35 ± 5.04
Avg Face	5.70 ± 1.11	7.43 ± 1.35	6.60 ± 1.12	5.48 ± 1.10	5.72 ± 1.08	6.15 ± 1.09	6.34 ± 1.08	5.90 ± 1.07	5.79 ± 1.13	5.43 ± 1.09
Avg	9.07 ± 1.28	10.25 ± 1.37	9.00 ± 1.21	7.64 ± 1.22	7.83 ± 1.22	7.74 ± 1.26	9.43 ± 1.30	7.53 ± 1.20	8.51 ± 1.28	8.49 ± 1.30

Landmark regions are as described in Salazar et al. (2014). *L* and *R* are shorthand for *Left* and *Right* respectively. *Avg Face* is the average for all inner face landmarks, and therefore excludes *Chin*, *L Face*, and *R Face*. No s.c. is short for no skip connections in the mesh convolutional decoder. Models indicated as “no m.m.” have no PCA mouth model or any form of regularization for the mouth region. S.d. stands for single decoder

Table 8 Semantic landmarks error on 3DMD for the ablations compared: Comparison of the mean and standard deviation for semantic landmark error (mm) for 3DMD using the *ibug* 68 facial landmark set

Region	SMF	SMF no s.c.	SMF fc	SMF fc’	SMF fc’ no m.m.	SMF fc lap=1e-3	SMF no m.m.	SMF lap=1e-3	SMF s.d.	SMF 512 s.d.
L Eyebrow	5.55 ± 1.76	5.28 ± 1.75	5.12 ± 1.53	6.45 ± 1.52	6.03 ± 1.58	5.49 ± 1.87	5.45 ± 1.44	4.87 ± 1.57	4.91 ± 1.44	6.19 ± 1.47
R Eyebrow	6.29 ± 2.32	5.93 ± 2.16	4.84 ± 1.79	6.57 ± 1.91	6.42 ± 1.83	6.09 ± 2.26	6.19 ± 1.93	4.79 ± 1.81	4.80 ± 1.77	5.67 ± 1.68
L Eye	4.26 ± 1.07	4.37 ± 1.10	3.69 ± 1.14	5.97 ± 1.41	6.13 ± 1.33	4.43 ± 1.07	3.99 ± 1.06	3.75 ± 1.14	3.67 ± 1.15	6.06 ± 1.44
R Eye	4.03 ± 1.31	3.76 ± 1.20	3.52 ± 1.26	5.75 ± 1.50	5.52 ± 1.39	4.38 ± 1.26	3.04 ± 1.11	3.58 ± 1.27	3.61 ± 1.31	6.28 ± 1.48
Nose	5.26 ± 0.87	4.77 ± 0.78	3.84 ± 0.81	5.66 ± 0.86	4.84 ± 0.81	5.29 ± 0.87	5.19 ± 0.86	4.24 ± 0.89	3.99 ± 0.82	5.25 ± 0.88
Mouth	6.31 ± 1.19	6.09 ± 1.23	4.87 ± 1.23	4.96 ± 1.32	5.10 ± 1.27	9.91 ± 1.24	6.71 ± 1.17	4.98 ± 1.23	5.88 ± 1.19	4.64 ± 1.29
Jaw	24.74 ± 5.95	25.23 ± 6.05	27.87 ± 5.77	27.34 ± 5.78	27.15 ± 5.77	24.57 ± 5.98	28.46 ± 5.75	27.91 ± 5.77	27.48 ± 5.78	26.68 ± 5.79
Avg Face	5.54 ± 0.85	5.29 ± 0.83	4.41 ± 0.81	5.60 ± 0.89	5.45 ± 0.85	7.00 ± 0.87	5.52 ± 0.81	4.51 ± 0.83	4.82 ± 0.79	5.36 ± 0.86
Avg	10.34 ± 1.92	10.27 ± 1.94	10.28 ± 1.83	11.04 ± 1.88	10.87 ± 1.86	11.39 ± 1.94	11.25 ± 1.84	10.36 ± 1.84	10.48 ± 1.83	10.69 ± 1.86

L and *R* are shorthand for *Left* and *Right* respectively. *Avg* is the average for all inner face landmarks. No s.c. is short for no skip connections in the mesh convolutional decoder. Models indicated as “no m.m.” have no PCA mouth model or any form of regularization for the mouth region. S.d. stands for single decoder

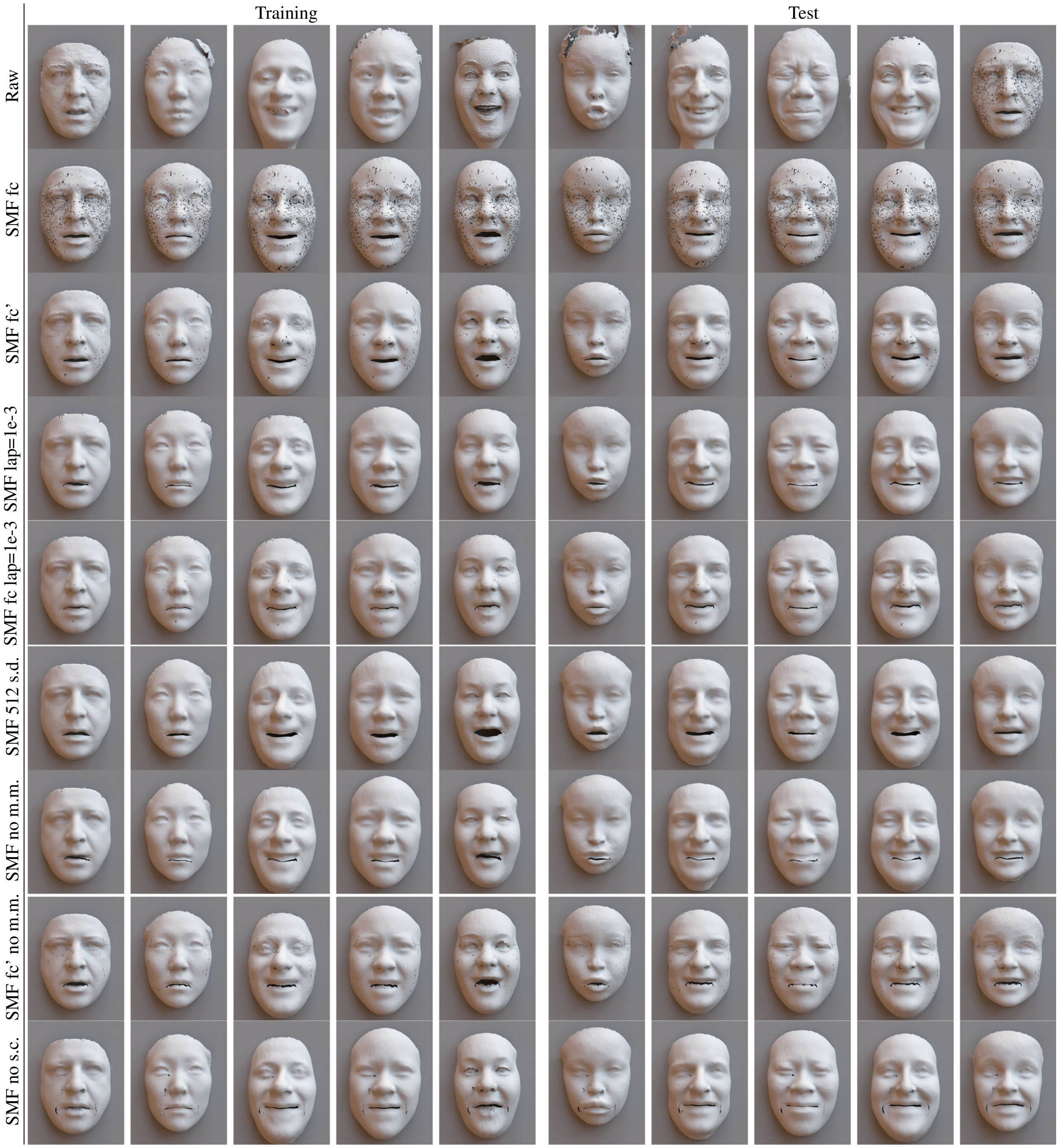


Fig. 20 Visual comparison of different ablations: We selected 5 training scans and 5 test scans from Figs. 14 and 15 and produced their registration with various choices of decoders compared in our ablation study

on BU3D and 3DMD in Table 7 and Table 8. We further provide sample reconstructions in Figure 20.

Numerically, the models with no mouth regularization showed higher test surface error and lower training surface error, for both fully-connected an mesh inception decoders, with the convolutional decoders markedly outperforming the dense layers. This is explained by the fact that not constraining the vertices in the mouth region enables the model to match them at a low cost (in terms of chamfer distance) with points from the teeth or inside of the mouth, thus lowering the error measured. Laplacian regularization behaves similarly, as visualized in Fig. 20, where the mouth reconstructions

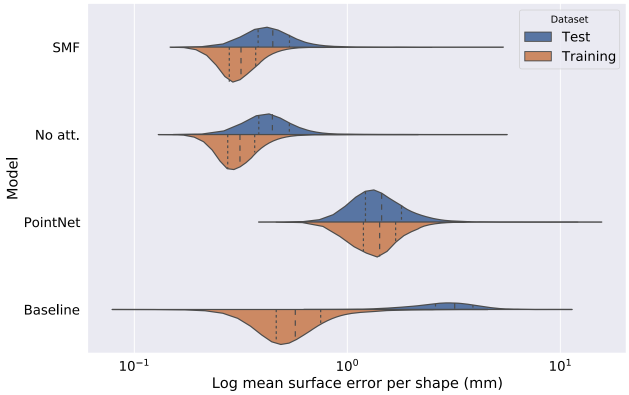


Fig. 21 Violin plot: of the training and test error for the model trained without attention compared to SMF, SMF with a vanilla PointNet encoder, and the baseline

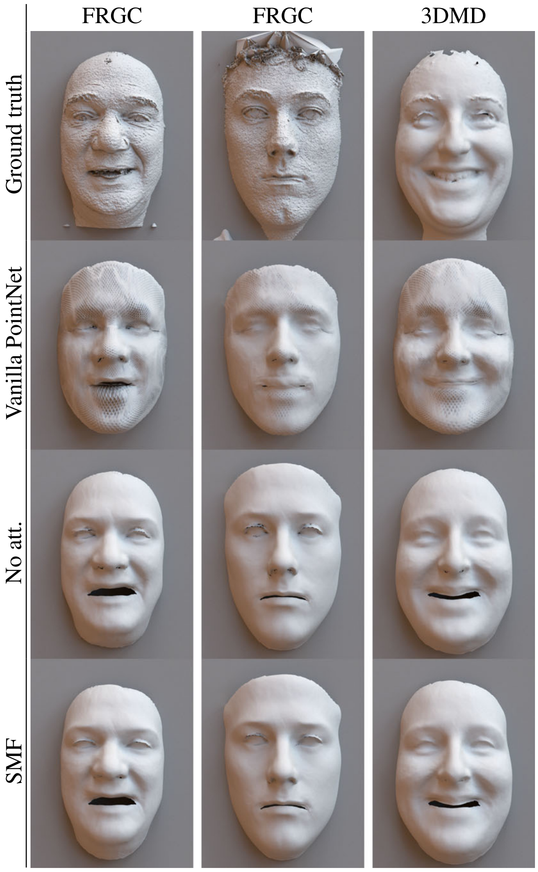


Fig. 22 Ablation study on the attention mechanism: The attention mechanism helps reduce noise and improve details on out of sample registrations

of the models that use Laplacian loss are in-between the non-regularized models and the PCA-neural network hybrids in terms of deformations induced by noisy points from the inside of the mouth. On the other hand, the hybrid models produced noise-free reconstructions in all cases.

We note that relying only on the neural network, with an additional Laplacian loss, improved the surface fairness of the registrations produced by the fully-connected decoders. This is expected, as the hybrid models ought to be harder to optimize. Naturally, the non-linear models are also more powerful and expressive than the PCA mouth models (which is the reason why we use the latter to constrain the former and perform denoising), and should therefore be favored when training on curated noise-free data.

We conclude that, for collections of raw noisy scans, our proposed approach of building hybrid models is effective.

5.5 Ablation Study on the Encoder

We now evaluate the contribution of the improvements we made to the PointNet encoder (attention mechanism, group normalization)by carrying-out an ablation study. We train SMF with our modified PointNet encoder without attention (No att.) and with a vanilla PointNet encoder. As a reminder, the baseline is evaluated on the processed (cropped, subdivided) data.

5.5.1 Distribution of the Surface Error

We visualize the distribution of the surface error on the 5000 training and test scans in Fig. 21, as well as that of the baseline.

As can be seen in Fig. 21, SMF with vanilla PointNet has lower training set performance than the baseline, which used a vanilla PointNet trained on cropped scans, but does not overfit contrary to the baseline. The distributions of the training error of SMF with and without attention are extremely close, with the no attention variant actually showing marginally lower error. As shown in Qi et al. (2017a), PointNet summarizes the input point cloud with a few (at most as many as the output dimension of the max pooling layer) points from the input. This property makes PointNet naturally robust to noise *to some extent*. When looking at the generalization gap for the models, we can see the surface error increased less for SMF than for the model without attention, as can be further verified in Fig. 19. These observations suggest our changes all contribute to improved performance and improved generalization. We verify the contribution of the attention mechanism visually in Fig. 22. We can see SMF without attention performs well, but reconstructions are noisier for the faceted scans from FRGC, and less details are present in the test 3DMD scan. Revisiting the examples of Fig. 2, we can also see the attention mechanism helps discard sensor noise in Fig. 23, and in line 3, col. 2 of Fig. 14, in which points inside the mouth also received attention scores close to 0.

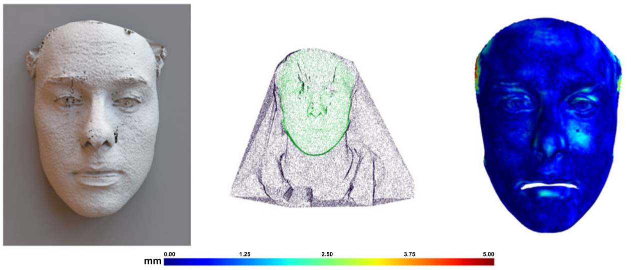


Fig. 23 Revisiting the example of Fig. 2: the attention mechanism is able to discard noisy points in badly-triangulated range scans

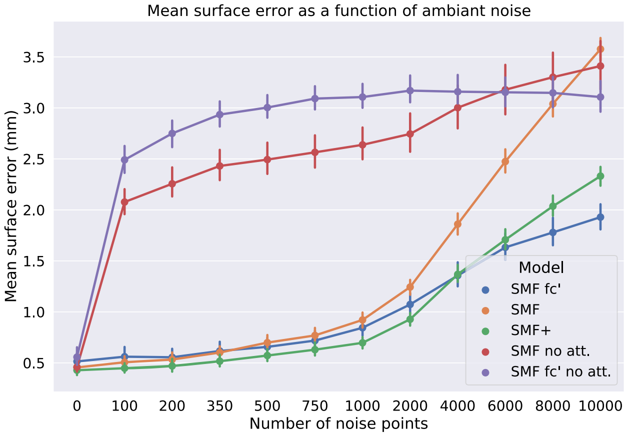


Fig. 24 Average mean surface error for increasing levels of noise measured on 100 randomly selected test scans from the 3DMD dataset. Models trained without our attention mechanism are very sensitive to random perturbations of their input, as shown by the sharp increase in mean surface error and the large variance of the surface error, even for low noise levels. Our models trained with attention are, on the other hand, more resilient to corruption

5.5.2 Ambient Noise

To better showcase the contribution of the attention mechanism, we now evaluate our trained models on 3DMD scans with additional artificial noise added. Our experimental setting is as follows: for a given 3DMD scan, we sample a set $\mathscr{P}$ of 2^{16} points at random on the scan. A second set $\mathscr{U}$ of N points is then drawn uniformly at random in an cubic volume containing the scan. Finally, our input point cloud $\mathbf{X}$ consists of 2^{16} points drawn uniformly at random and without replacement from $\mathscr{P} \cup \mathscr{U}$. Examples of resulting point clouds are shown in Fig. 25 for $N = 500$.

We apply SMF, SMF+, SMF fc’ as well as SMF and SMF fc’ without attention to $\mathbf{X}$ and measure the mean surface error between the reconstructions and the raw scan. In total, we repeated this procedure for 100 scans and 11 noise levels ranging from no noise ($N = 0$) to substantial noise ($N = 10{,}000$). We report the results in Fig. 24.

As can be seen from Fig. 24, the models that do not have an attention mechanism are very sensitive to noise. As little as 100 random points prior to sampling the input point cloud lead to significant deformations of the output, regardless of the choice of decoder. This is apparent when visualizing the registrations, e.g. for a test subject from the 3DMD dataset in Fig. 26. On the other hand, the models trained with attention are more robust: the surface error increases slower, and has lower variance as indicated by the shorter error bars. Visually, the reconstructions we obtain from the noisy inputs are indistinguishable from the noise-free inputs for low noise levels. We note, however, that not all models with attention learn equally good segmentations of the input point cloud. In this particular case, our SMF model was more susceptible to noise than SMF+ and SMF fc. We compare the attention masks of the noisy point clouds of some models in Fig. 25, and verify that the segmentations isolate the most relevant points.

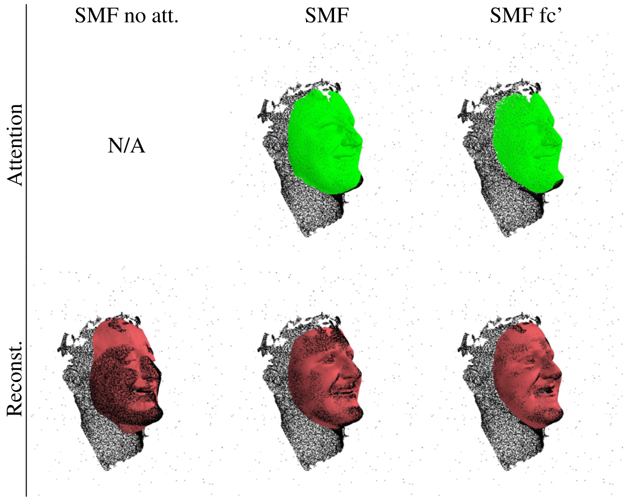


Fig. 25 Noisy input, attention mask (green) and registration for three models trained with attention, and 500 noisy points added prior to sampling the input point cloud. Clear segmentations are obtained in all three cases, with noise points receiving a low attention score even for points close to the actual scan. This results in markedly more robust registrations

We will further verify that the attention mechanism improves the quality of reconstructions on noisy out of sample scans in Sect. 6.4.

5.6 Overview

In Sects. 5.1 and 5.2, we showed SMF (and SMF+) systematically outperforms the baseline on landmark localization error and offers performance competitive with NICP. Test set performance, in particular, was markedly higher than the current state of the art, and remained very close to the training set error. Direct evaluation of the mean surface registration error offers a more complete picture of the registration quality and leads to similar conclusions. Visual inspection of the reconstructions confirms the quantitative analysis: contrary

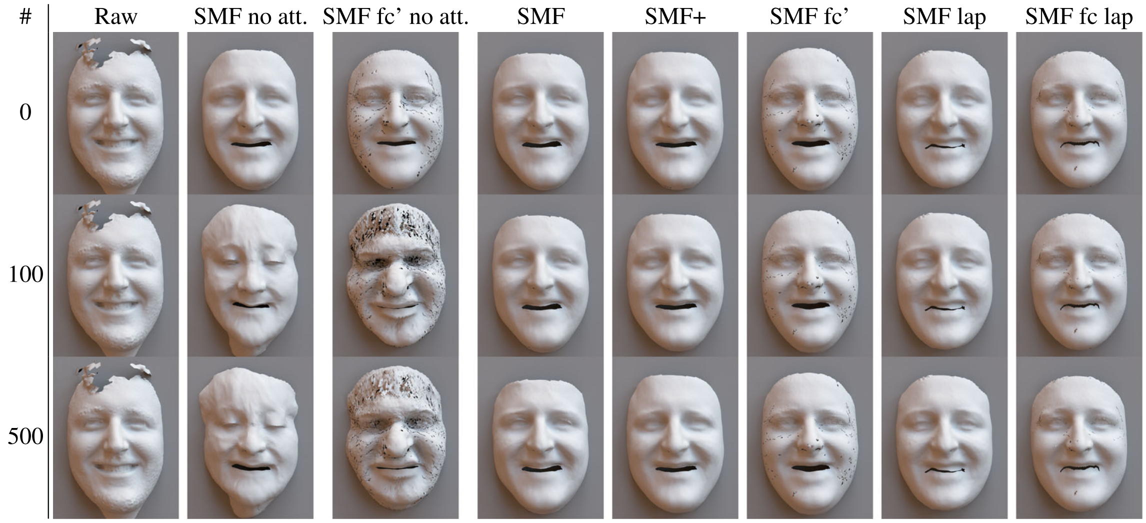


Fig. 26 Sample registrations for artificial ambient noise for several choices of decoders and mouth regularization. $\lambda_{edge} = 2e-4$ for SMF fc' and SMF fc' no att., and $\lambda_{edge} = 5e-5$ for SMF fc lap. We set $\lambda_{lap} = 1e-3$

to the baseline, SMF provides noise-free registrations which closely match the raw scans in both identity and expression. We showed re-sampling the scans typically lead to minor variations in their registrations in Sect. 5.3. In Sect. 5.4, we compared our default architecture of two mesh inception convolutional decoders and PCA mouth models with different variations, namely fully-connected decoders, using a single decoder, not using skip connections, not using any mouth regularization, or using uniform Laplacian regularization of the mouth region. We showed our contributions provide tangible benefits in reconstruction accuracy and robustness for noisy raw scans data, while our framework is flexible enough to accommodate various substitutions while preserving the ability of the models to generalize well to unseen data. Finally, we evaluated the contributions of our modifications of the PointNet encoder in Sect. 5.5. In particular, we demonstrated that our attention mechanism markedly improves the models' robustness to random perturbations of their input in the form of ambient noise, regardless of our choice of decoder. This demonstrates that our contributions to the encoding and decoding stages are both orthogonal and complementary.

6 A Large Scale Hybrid Morphable Model

In this section, we assess the morphable model aspects of SMF. We first study the influence of the dimension of the identity and expression latent spaces on surface reconstruction error both in sample and out of sample. We then show SMF can be used to quickly generate realistic synthetic faces. In Sect. 6.3, we evaluate SMF on shape-to-shape translation applications, namely identity and expression tranfer, and morphing. We conclude by showing SMF can be used successfully for registration and translation fully in the wild.

6.1 Dimension of the Latent Spaces

The classical linear morphable models literature typically reports three main metrics. *Specificity* is evaluated in Sect. 6.2.1. *Compactness* is the proportion of the variance retained for increasing numbers of principal components—a direct correlate of the training error for PCA models. *Generalization* measures the reconstruction error on the test set for increasing numbers of principal components. Since our model is not linear, we instead report the training and test performance for increasing identity and expression dimensions. We choose symmetric decoders with $\mathbf{z}_{id}$ and $\mathbf{z}_{exp}$ of equal dimension d. We vary $d \in \{64, 128, 256, 512\}$. We measure the mean (per scan) surface reconstruction error on the same subsets of 5000 training and 5000 test scans used in Sect. 5. We plot the mean error across the 5000 scans along with its 95% confidence interval obtained by bootstrapping in Fig. 27.

As expected, both the training and test error decrease steadily up to $d = 256$. For $d = 512$, our data shows increased training and test error compared to $d = 256$. This shows there is diminishing return in increasing the model complexity, and bolsters our choice of $d = 256$ for SMF.

6.2 Generating Synthetic Faces

We now evaluate the generative ability of our SMF+ model.

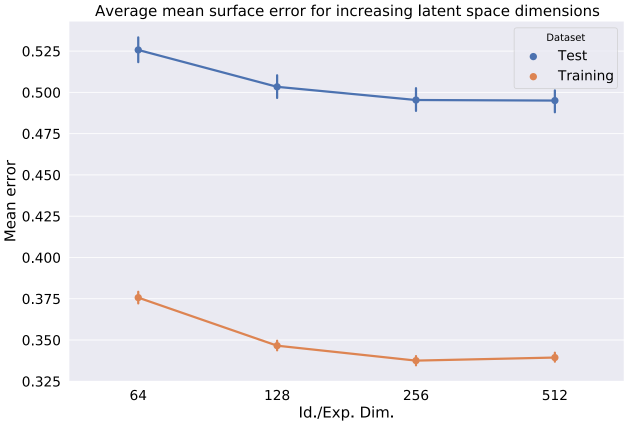


Fig. 27 Compactness and generalisation: Training and test error for increasing number of latent dimensions

6.2.1 Specificity Error

We follow the literature and measure the specificity error as follows: we sample 10,000 shapes at random from the joint latent space; since our model is not explicitly trained as a generative model, no particular structure is to be expected on the latent space and we therefore model the empirical distribution of the joint latent vectors of the training set with a multivariate Gaussian distribution. We estimate the empirical mean and covariance matrix of the $\approx$ 54,000 joint latent vectors and generate 10,000 Gaussian random vectors. We apply the pre-trained decoder to obtain generated faces.[1]

For each of the 10,000 random faces, we find its closest point in the training set in terms of minimum (over all 54,000 training registrations) of the average (over the 29,495 points in the template) vertex-to-vertex Euclidean distance. The mean of these 10,000 distances is the specificity error of the model. For the sake of completeness, we repeated the experiment with the variants of SMF evaluated in Fig. 27. We plot the specificity error and its 95% confidence interval computed by bootstrapping in Fig. 28. Both SMF and SMF+ offer low specificity error, suggesting realistic-looking samples can be obtained. SMF+, in particular, has markedly lower specificity error than SMF for the same latent space dimensions, which confirms the benefits of training our very large scale model on the extended training set.

6.2.2 Visualization of the Samples

We now inspect a random subset of the 10,000 samples in Fig. 29. We render each random sample, its closest point in the registered training set, and the raw scan from which the registration was obtained. We can see the samples generated by SMF+ are highly diverse and realistic-looking: they are close to the registrations of the training set without displaying mode collapse. SMF+ generates detailed faces with sharp features across a wide range of identity, age, ethnic background, and expression, including extreme face and mouth expressions. We further note the absence of artifacts and the seamless blending of the mouth with the rest of the face.

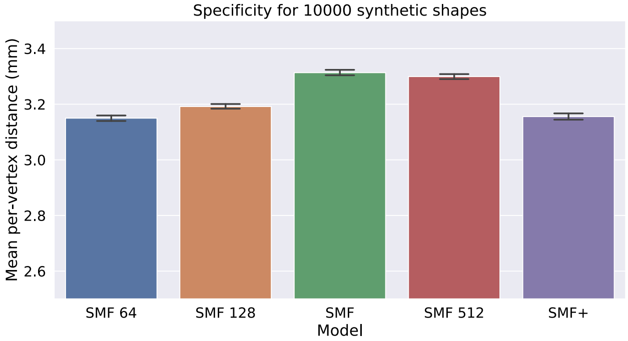


Fig. 28 Specificity error: for variants of SMF and SMF+. The specificity error is the mean distance of the sampled scans to their projection on the registered training set

6.3 Interpolation in the Latent Space

We now present a surface-to-surface translation experiment on the training set by showing the results of expression transfer and identity and expression interpolation in the latent spaces of SMF+. Since the latent vectors are hyperspherical, care must be taken to interpolate along the geodesics on the manifold. We therefore interpolate between two latent vectors $\mathbf{z}_1$ and $\mathbf{z}_2$ and $t \in [0, 1]$ as

$$\mathbf{z}_i = \frac{\mathbf{z}_1 + t(\mathbf{z}_1 - \mathbf{z}_2)}{||\mathbf{z}_1 + t(\mathbf{z}_2 - \mathbf{z}_1)||_2}. \qquad (23)$$

We select two expressive scans of two different subjects, referred to as S1 and S2, from two different databases (BU-3DFE and BU-4DFE) displaying distinct expressions (disgust and surprise). We study three cases: simultaneous interpolation of identity and expression, interpolation of identity for a fixed expression, and interpolation of expression for a fixed identity. We render points along the trajectory defined by Equation 23 at $t \in \{0, 0.25, 0.5, 0.75, 1\}$. The results of the interpolation are presented in Fig. 30.

We observe smooth interpolation in all three cases. For simultaneous interpolation, we obtain a continuous morphing of the first expressive scan ($t = 0$) into the second expressive scan ($t = 1$). In particular, we note that the midpoint resembles what would be the neutral scan of a subject presenting physical traits of both the source (nose, forehead) and destination (eyes, jawline) subjects. The interpolation of the identity vector for the fixed expression of S1 shows a smooth

[1] Generating all 10,000 random faces took 55s on a single consumer-grade GPU.

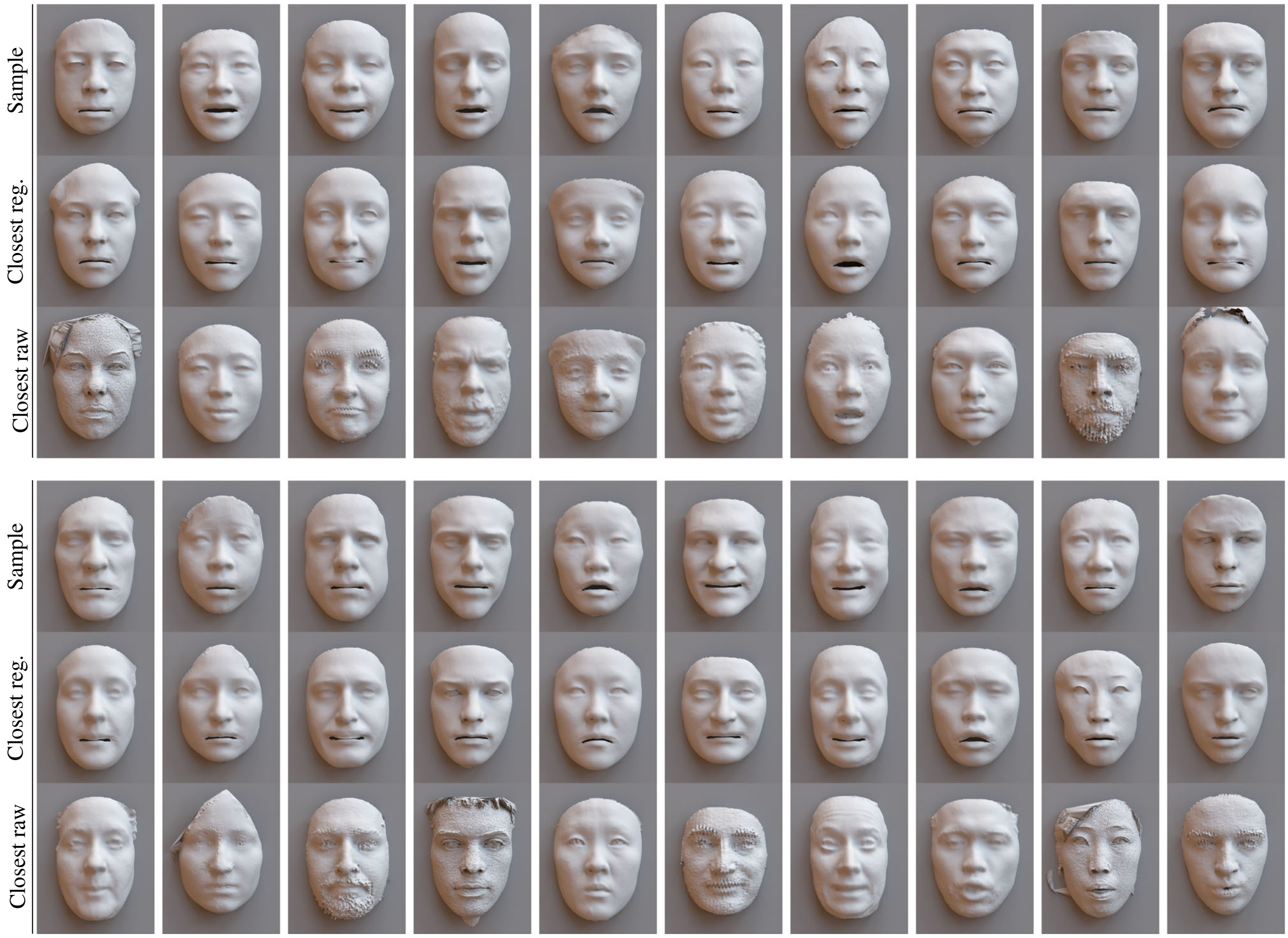


Fig. 29 Samples from SMF+: First row: Generated face obtained by sampling a random joint vector. Second row: Closest registration in the training set. Third row: Raw scan from which the closest registration was obtained

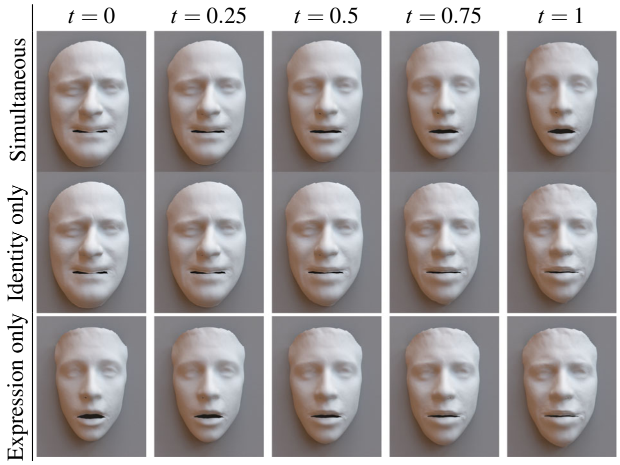


Fig. 30 Interpolation on the training set: joint interpolation of identity and expression, and interpolation over one factor with the other factor fixed

transition towards S2 while keeping the correct expression. Conversely, interpolation between S2 and S2 with the expression of S1 shows the overall identity is recognizable and the expression displays a smooth evolution from surprise to disgust. These results show our model can be used for expression transfer and smooth interpolation on the training set. In Sect. 6.4, we evaluate SMF on surface-to-surface translation tasks in the wild.

6.4 Face Modeling and Registration in the Wild

We now evaluate SMF on the difficult tasks of registration and manipulation of scans found "in the wild", i.e. in uncontrolled environments, with arbitrary sensor types and acquisition noise. We collected the scans of three subjects, referred to A, B, and C, in various conditions. For subject A (male, Caucasian), we obtained crops of two body scans, acquired at over a year and half's interval using two different body capture setups that produce meshes, in two different environments. The first scan shows a crop of the subject squatting while raising his right eyebrow, the second is of the subject jumping with a neutral face. We further acquired four high

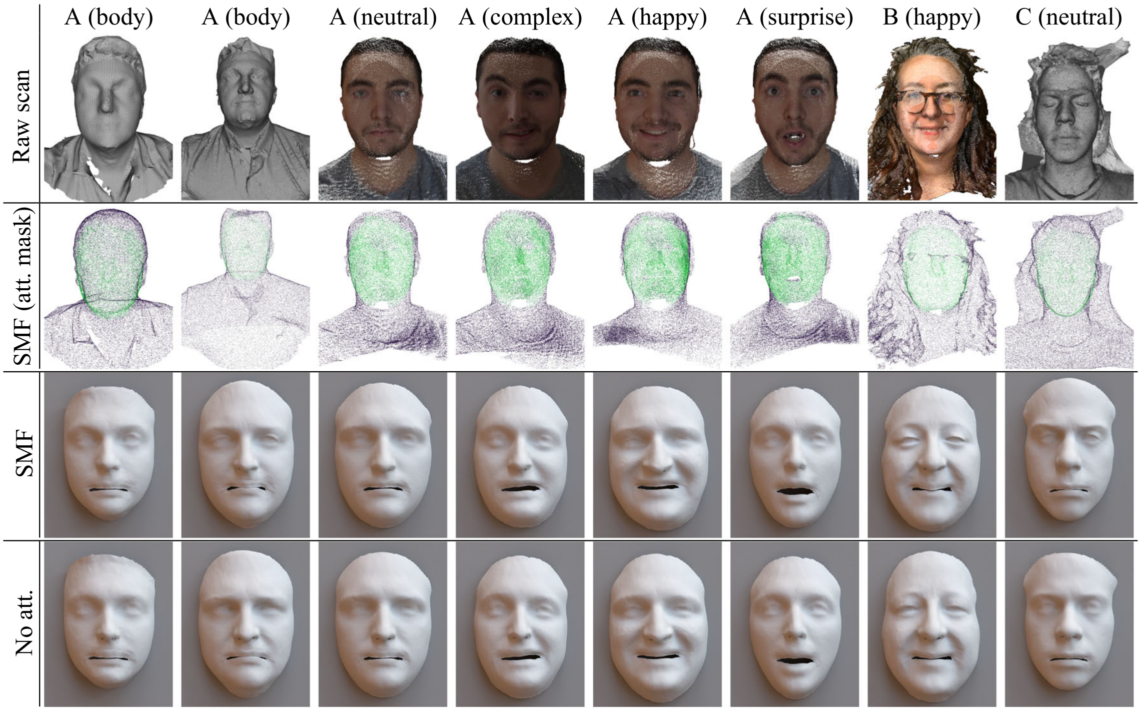


Fig. 31 In the wild registrations with and without attention: the scans of subject A were acquired over a period of two years using three different cameras (two different body capture stages and a commodity depth sensor in a smartphone). The scan of subject B was also acquired using a smartphone depth camera, but using a lower resolution setting. The scan of subject C is from a state of the art facial scanning light stage. SMF provides consistent high-quality registrations even from low-resolution scans comprising large areas of the body, hair, or background. In particular, the six scans of subject A show consistent representation of the identity. The attention mechanism can be seen to improve details in the registrations

density point clouds of subject A performing different facial expressions : neutral, smiling (happy), surprise, and a "complex" compound expression consisting of raising the right eyebrow while opening and twisting the mouth to the left. Scanning was done in an uncontrolled environment using a commodity sensor, namely the embedded depth camera of an iPhone 11 Pro. Subject B (female, Caucasian) was captured posing with a light smile in a different uncontrolled environment, also with an iPhone 11 Pro, but using a lower resolution point cloud. Finally, subject C (male, Caucasian) was captured in a neutral pose using a state of the art light stage setup that outputs very high resolution meshes. All in all, the scans represent four different cameras, in five different environments, at five different levels of detail and surface quality, and across two different modalities (mesh and point cloud).

We use the pre-trained SMF model with and without attention to further extend the ablation study of Sect. 5.5. Scans were rigidly aligned with the cropped LSFM mean using landmarks. For meshes (body scans, light stage scan), we sample 2^{16} input points at random on the surface of the triangular mesh. For point clouds, we select 2^{16} points.

Figure 31 shows the raw scans, registration from SMF, predicted attention mask for SMF, and registration for SMF trained without visual attention. We can see SMF produced very consistent registrations for subject A across modalities, resolution, and time: it is clear, from the registrations, that the scans came from the same subject, even for the low-resolution face and shoulders region of the first body scan, for which important facial features and the elevated position of the right eyebrow were captured. Comparing the neutral iPhone scan and the neutral body scan further shows identity was robustly captured at the two different resolutions. The highly non-linear complex expression was, also, accurately captured, and so were the more standard happy and surprise expressions. Performance was stable for lower-resolution raw point clouds too as shown with the registration of subject B. SMF produced a sharp detailed registration of the high quality light-stage scan of subject C, correctly capturing the shape of the nose, the sharpness and inflexion of the eyebrows, and the angle of the mouth.

Compared to SMF, SMF trained without our attention mechanism still produced high quality registrations but with fewer details. The two body scans and the light stage scans show clear differences, especially in the eyes. The happy expression of subject B was not captured as accurately, and the shape of the face appears elongated. Looking at the attention masks, we can see our visual attention mechanism

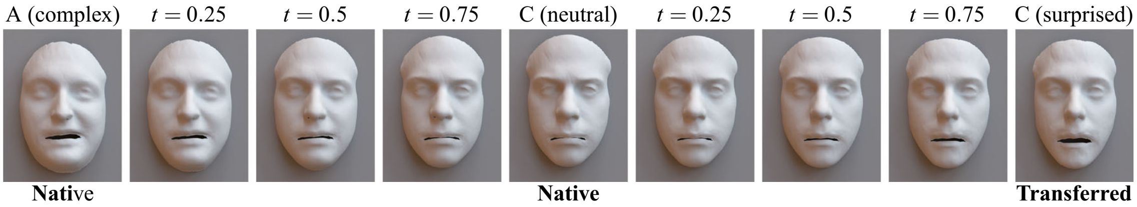


Fig. 32 Interpolation, transfer, and morphing in the wild: From A "complex" to C "neutral" to C "surprised" transferred from A

discarded points from the body, the inside of the mouth (A surprise), environment noise (C neutral), and hair and partial occlusions (B happy, for which it removed most of the glasses).

Morphing and editing in the wild We now show our pre-trained model can be used for shape morphing and editing, such as expression transfer, by linearly interpolating in $\mathscr{S}^{255}$ between the predicted identity and expression vectors of the raw scans. We select the 'A complex", "A surprise" and "C neutral" scans and register both of them with our pre-trained SMF model, keeping their predicted identity and expression embeddings. We first interpolate the identity and expression jointly between "A complex" and "C neutral" to produce a smooth morphing. We then keep the identity vector fixed to that of "C neutral" and linearly interpolate between the expression vectors of "C neutral" and "A surprise", this produces a smooth expression transfer. Both experiments are shown as a continuous transformation in Fig. 32.

As apparent from Fig. 32, our model is able to smoothly interpolate between subjects and expressions of scans captured, in the wild, across different modalities and resolutions. The morphing from A complex to C neutral produces smooth facial motions without discontinuities. Our model is further able to, not only transfer expressions in the wild, but smoothly interpolate between expression vectors of different subjects for a fixed identity. The interpolation transfer again produces a smooth natural-looking transition between the neutral scan of C, with the mouth and eyebrows smoothly moving from a resting position to a surprise expression, while keeping the facial features of subject C.

7 Conclusion and Future Work

In this paper, we present Shape My Face (SMF), a novel learning-based algorithm that treats the registration task as a surface-to-surface translation problem. Our model is based on an improved point cloud encoder made highly robust with a novel visual attention mechanism, and on our mesh inception decoders that leverage graph convolutions to learn a compact non-linear morphable model of the human face. We further improve robustness to noise in face scans by blending the output of the mesh convolutions with a specialized statistical model of the mouth in a seamless way. Our model learns to produce high quality registrations both in sample and out of sample, thanks to the improved weight sharing and stochastic training approach that prevent the model from overfitting any particular discretization of the training scans.

We introduce a large scale morphable model, coined as SMF+, by training SMF on 9 comprehensive human 3D facial databases. Our experimental evaluation shows SMF+ can generate thousands of diverse realistic-looking faces from random noise across a wide range of age, ethnicities, genders, and (extreme) facial expressions. We evaluate SMF+ on shape editing and translation tasks and show our model can be used for identity and expression transfer and interpolation. Finally, we show SMF can also accurately register and interpolate between facial scans captured in uncontrolled conditions for unseen subjects and sensors, allowing for shape editing entirely in the wild. In particular, we demonstrated smooth interpolation and transfer of expression and identity between a very high quality mesh acquired in controlled conditions with a sophisticated facial capture environment, and a noisy point cloud produced by consumer-grade electronics.

Future work will investigate improving the reproduction of high frequency details in the scans, and registering texture and geometry simultaneously.

Funding M. Bahri was supported by a Department of Computing scholarship and a Qualcomm Innovation Fellowship. E. O' Sullivan was supported by a Department of Computing scholarship. S. Zafeiriou was partially funded by the EPSRC Fellowship DEFORM: Large Scale Shape Analysis of Deformable Models of Humans (EP/S010203/1). S. Gong and M. M. Bronstein were partially funded by ERC Consolidator grant No. 724228 (LEMAN). We thank Amazon for AWS Cloud Credits for Research.

Availability of Data and Materials All databases of Table 1 are available upon request to their respective authors and sufficient for reproducing the main results of the paper. 4DFAB and MeIn3D (used to train SMF+) and 3DMD (used for testing) are not publicly available currently.

Declarations

Conflict of interest The authors declare no conflicts of interest.

Code availability A pre-trained model will be released publicly along with code, please visit https://github.com/mbahri/smf.

References

Abrevaya, V. F., Wuhrer, S., & Boyer, E. (2018). Multilinear autoencoder for 3D face model learning. In *Proceedings—2018 IEEE winter conference on applications of computer vision, WACV 2018* (Vol. 2018, pp. 1–9).

Amberg, B., Romdhani, S., & Vetter, T. (2007). Optimal step nonrigid ICP algorithms for surface registration. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, IEEE (pp. 1–8).

Amberg, B., Knothe, R., & Vetter, T. (2008). Expression invariant 3D face recognition with a morphable model. In *2008 8th IEEE international conference on automatic face and gesture recognition, FG 2008*.

Aoki, Y., Goforth, H., Srivatsan, R. A., & Lucey, S. (2019). Pointnetlk: Robust & efficient point cloud registration using pointnet. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (Vol. 2019, pp. 7156–7165).

Aytekin, C., Ni, X., Cricri, F., & Aksu, E. (2018). Clustering and unsupervised anomaly detection with l2 normalized deep auto-encoder representations. In *Proceedings of the international joint conference on neural networks*.

Bagautdinov, T., Wu, C., Saragih, J., Fua, P., & Sheikh, Y. (2018). Modeling facial geometry using compositional VAEs. In *The IEEE conference on computer vision and pattern recognition (CVPR)*.

Bagdanov, A. D., Masi, I., & Del Bimbo, A. (2011). The florence 2D/3D hybrid face datset. In *Proceedings of ACM multimedia internationl workshop on multimedia access to 3D human objects (MA3HO'11)*. ACM, ACM Press.

Baocai, Y., Yanfeng, S., Chengzhang, W., & Yun, G. (2009). BJUT-3D large scale 3D face database and information processing. *Journal of Computer Research and Development*, *46*(6), 1009.

Besl, P. J., & McKay, N. D. (1992). A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *14*(2), 239–256.

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on computer graphics and interactive techniques, SIGGRAPH 1999*.

Blanz, V., & Vetter, T. (2003). Face recognition based on fitting a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine*, *25*, 1063–1074. Intelligence.

Bolkart, T., & Wuhrer, S. (2015). A groupwise multilinear correspondence optimization for 3D faces. In *Proceedings of the IEEE international conference on computer vision*.

Bookstein, F. L. (1989). Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*, 567–585.

Booth, J., Roussos, A., Zafeiriou, S., Ponniahy, A., & Dunaway, D. (2016). A 3D morphable model learnt from 10,000 faces. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE (pp. 5543–5552).

Booth, J., Antonakos, E., Ploumpis, S., Trigeorgis, G., Panagakis, Y., & Zafeiriou, S. (2017). 3D face morphable models "In-the-Wild". In *Proceedings—30th IEEE conference on computer vision and pattern recognition, CVPR 2017*.

Booth, J., Roussos, A., Ponniah, A., Dunaway, D., & Zafeiriou, S. (2018a). Large scale 3D morphable models. *International Journal of Computer Vision*, *126*(2–4), 233–254.

Booth, J., Roussos, A., Ververas, E., Antonakos, E., Ploumpis, S., Panagakis, Y., et al. (2018b). 3d reconstruction of "in-the-wild" faces in images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(11), 2638–2652.

Boscaini, D., Masci, J., Rodolá, E., & Bronstein, M. (2016). Learning shape correspondence with anisotropic convolutional neural networks. In *Proceedings of the 30th international conference on neural information processing systems* (pp. 3197–3205).

Bouritsas, G., Bokhnyak, S., Ploumpis, S., Bronstein, M., & Zafeiriou, S. (2019). Neural 3D morphable models: Spiral convolutional networks for 3D shape representation learning and generation. In *The IEEE international conference on computer vision (ICCV)*.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., & Vandergheynst, P. (2017). Geometric deep learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine*, *34*(4), 18–42.

Bruna, J., Zaremba, W., Szlam, A., & LeCun, Y. (2014). Spectral networks and deep locally connected networks on graphs. In *2nd international conference on learning representations, ICLR 2014—Conference track proceedings* (pp. 1–14).

Burt, P. J., & Adelsonm, E. H., (1985). Merging images through pattern decomposition. In *Applications of digital image processing VIII*.

Cao, C., Weng, Y., Zhou, S., Tong, Y., & Zhou, K. (2014). FaceWarehouse: A 3D facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, *20*, 413–425.

Chen, Y., & Medioni, G. (1991). Object modeling by registration of multiple range images. In *Proceedings—IEEE international conference on robotics and automation* (Vol. 3, pp. 2724–2729).

Cheng, S., Marras, I., Zafeiriou, S., & Pantic, M. (2017). Statistical non-rigid ICP algorithm and its application to 3d face alignment. *Image and Vision Computing*, *58*, 3–12.

Cheng, S., Kotsia, I., Pantic, M., & Zafeiriou, S. (2018). 4DFAB: A large scale 4D database for facial expression analysis and biometric applications. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 5117–5126).

Clevert, D. A., Unterthiner, T., & Hochreiter, S. (2016). Fast and accurate deep network learning by exponential linear units (ELUs). In *4th international conference on learning representations, ICLR 2016—Conference track proceedings*.

Crane, K., Weischedel, C., & Wardetzky, M. (2017). The heat method for distance computation. *Communications of the ACM*, *60*(11), 90–99.

Crane, K., Vaz, C., & Fabri, A. (2020). The heat method. In *CGAL user and reference manual* (5th ed.). CGAL Editorial Board

Dai, H., Pears, N., Smith, W., & Duncan, C. (2017). A 3D morphable model of craniofacial shape and texture variation. In *Proceedings of the IEEE international conference on computer vision*.

De Smet, M., & Van Gool, L. (2011). Optimal regions for linear model-based 3D face reconstruction. In *Lecture notes in computer science (including subseries Lecture notes in artificial intelligence and Lecture notes in bioinformatics)*.

Defferrard, M., Bresson, X., & Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29, pp. 3844–3852). Curran Associates Inc.

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*.

Egger, B., Smith, W. A. P., Tewari, A., Wuhrer, S., Zollhoefer, M., Beeler, T., et al. (2020). 3D morphable face models-past, present, and future. *ACM Transactions on Graphics*, *39*(5), 1–38. https://doi.org/10.1145/3395208.

Fey, M., Lenssen, J. E., Weichert, F., & Muller, H. (2018). SplineCNN: Fast geometric deep learning with continuous B-spline kernels. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (pp. 869–877)

Feydy, J., Charlier, B., Vialard, F. X., & Peyré, G. (2017). Optimal transport for diffeomorphic registration. In: *Lecture notes in computer science (including subseries Lecture notes in artificial intelligence and Lecture notes in bioinformatics)*.

Garland, M., & Heckbert, PS. (1997). Surface simplification using quadric error metrics. In *Proceedings of the 24th annual conference on computer graphics and interactive techniques. SIGGRAPH '97* (pp. 209–216). ACM Press/Addison-Wesley Publishing Co.

Gerig, T., Morel-Forster, A., Blumer, C., Egger, B., Luthi, M., Schoenborn, S., & Vetter, T. (2018). Morphable face models—An open framework. In *2018 13th IEEE international conference on automatic face gesture recognition (FG 2018)* (pp. 75–82). https://doi.org/10.1109/FG.2018.00021.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O.,& Dahl, G. E. (2017). Neural message passing for quantum chemistry. In *34th international conference on machine learning, ICML 2017* (Vol. 3, pp. 2053–2070).

Gong, S., Chen, L., Bronsteinm M., & Zafeiriou, S. (2019). SpiralNet++: A fast and highly efficient mesh convolution operator. In *The IEEE international conference on computer vision (ICCV) workshops*.

Gong, S., Bahri, M., Bronstein, MM., & Zafeiriou, S. (2020). Geometrically principled connections in graph neural networks. In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Gupta, S., Castleman, K. R., Markey, M. K., & Bovik, A. C. (2010). Texas 3D face recognition database. In *Proceedings of the IEEE southwest symposium on image analysis and interpretation*. IEEE (pp. 97–100).

Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. In *Advances in neural information processing systems* (Vol. 2017, pp. 1025–1035).

He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. In *2017 IEEE international conference on computer vision (ICCV)*. IEEE (pp. 2980–2988).

Horn, B. K., & Schunck, B. G. (1981). Determining optical flow. *Artificial Intelligence*, *17*(1–3), 185–203.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417–441.

Joo, H., Simon, T., & Sheikh, Y. (2018). Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*.

van Kaick, O., Zhang, H., Hamarneh, G., & Cohen-Or, D. (2011). A survey on shape correspondence. *Eurographics Symposium on Geometry Processing*, *30*(6), 1681–1707.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *Pattern Recognition Letters*, *94*, 172–179.

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional neural networks. *ICLR*, *2017*, 1–14.

Knoops, P. G., Papaioannou, A., Borghi, A., Breakey, R. W., Wilson, A. T., Jeelani, O., et al. (2019). A machine learning framework for automated diagnosis and computer-assisted planning in plastic and reconstructive surgery. *Scientific Reports*, *9*(1), 1–12.

Kolotouros, N., Pavlakos, G., Black, M., & Daniilidis, K. (2019a). Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE international conference on computer vision* (Vol. 2019, pp. 2252–2261).

Kolotouros, N., Pavlakos, G., & Daniilidis, K. (2019b). Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition* (Vol. 2019, pp. 4496–4505).

Lefébure, M., & Cohen, L. D. (2001). Image registration, optical flow, and local rigidity. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2106*, 26–38.

Lei, H., Akhtar, N., & Mian, A. (2019). Octree guided CNN with spherical kernels for 3d point clouds. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 9623–9632).

Li, G., Muller, M., Thabet, A., & Ghanem, B. (2019). DeepGCNs: Can GCNs go as deep as CNNs? In *The IEEE international conference on computer vision (ICCV)*.

Li, J., & Zhang, C. (2019). *Iterative matching point*. arXiv

Li, Q., Han, Z., & Wu, X. M. (2018). Deeper insights into graph convolutional networks for semi-supervised learning. In *32nd AAAI conference on artificial intelligence, AAAI 2018*.

Li, T., Bolkart, T., Black, M. J., Li, H., & Romero, J. (2017). Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics*, *36*, 194-1.

Lim, I., Dielen, A., Campen, M., & Kobbelt, L. (2018). A simple approach to intrinsic correspondence learning on unstructured 3D meshes. In *Computer vision—ECCV 2018 workshops—Munich, Germany, September 8–14, 2018, Proceedings, Part III* (pp. 349–362).

Liu, F., Tran, L., & Liu, X. (2019). 3D face modeling from diverse raw scan data. In *The IEEE international conference on computer vision (ICCV)*.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., & Black, M. J. (2015). SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, *34*, 1–16.

Lu, W., Wan, G., Zhou, Y., Fu, X., Yuan, P., & Song, S. (2019). DeepVCP: An end-to-end deep neural network for point cloud registration. In *Proceedings of the IEEE international conference on computer vision* (Vol. 2019, pp. 12–21).

Lucas, B. D., & Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th international joint conference on artificial intelligence (IJCAI)* (Vol. 2, pp. 674–679).

Luthi, M., Gerig, T., Jud, C., & Vetter, T. (2018). Gaussian process morphable models. *IEEE Transactions on Pattern Analysis and Machine*, *40*, 1860–1873. Intelligence.

Lüthi, M., Gerig, T., Jud, C., & Vetter, T. (2018). Gaussian process morphable models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(8), 1860–1873.

Masci, J., Boscaini, D., Bronstein, M. M., & Vandergheynst, P. (2015). Geodesic convolutional neural networks on Riemannian manifolds. In *Proceedings of the IEEE international conference on computer vision* (Vol. 2015, pp. 832–840).

Monti, F., Boscaini, D., Masci, J., Rodolá, E., Svoboda, J., & Bronstein, M. M. (2017). Geometric deep learning on graphs and manifolds using mixture model CNNs. In *Proceedings—30th IEEE conference on computer vision and pattern recognition, CVPR 2017* (Vol. 2017, pp. 5425–5434).

Mueller, A., Paysan, P., Schumacher, R., Zeilhofer, H. F., Berg-Boerner, B. I., Maurer, J., et al. (2011). Missing facial parts computed by a morphable model and transferred directly to a polyamide laser-sintered prosthesis: An innovation study. *British Journal of Oral and Maxillofacial Surgery*, *49*(8), e67–e71.

Myronenko, A., & Song, X. (2010). Point set registration: Coherent point drifts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*, 2262–2275.

Myronenko, A., Song, X., & Carreira-Perpiñán, M. Á. (2007). Non-rigid point set registration: Coherent point drift. *Advances in Neural Information Processing Systems*, *19*, 1009.

Nimier-David, M., Vicini, D., Zeltner, T., & Jakob, W. (2019). Mitsuba 2: A retargetable forward and inverse renderer. *Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, *38*(6), 1–17.

Patel, A., & Smith, W. A. P. (2009). 3D morphable face models revisited. In *2009 IEEE conference on computer vision and pattern recognition*. IEEE (pp. 1327–1334).

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, *2*(6), 559–572.

Pharr, M., Jakob, W., & Humphreys, G. (2016). *Physically based rendering: From theory to implementation* (3rd ed.). Morgan Kaufmann Publishers Inc.

Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Hoffman, K., Marques, J., Min, J., & Worek, W. (2005). Overview of the face recognition grand challenge. In *Proceedings—2005 IEEE computer society conference on computer vision and pattern recognition, CVPR 2005*.

Ploumpis, S., Wang, H., Pears, N., Smith, W. A., & Zafeiriou, S. (2019). Combining 3D morphable models: A large scale face-and-head model. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*.

Ploumpis, S., Ververas, E., O'Sullivan, E., Moschoglou, S., Wang, H., Pears, N., et al. (2020). Towards a complete 3D morphable model of the human head. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,. https://doi.org/10.1109/TPAMI.2020.2991150.

Qi, C. R., Su, H., Kaichun, M., & Guibas, L. J. (2017a). PointNet: Deep learning on point sets for 3D classification and segmentation. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE (pp. 77–85).

Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017b). PointNet++: Deep hierarchical feature learning on point sets in a metric space. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30, pp. 5099–5108). Curran Associates Inc.

Ranjan, A., Bolkart, T., Sanyal, S., & Black, M. J. (2018). Generating 3D faces using convolutional mesh autoencoders. In *The European conference on computer vision (ECCV)*.

Romero, J., Tzionas, D., & Black, M. J. (2017). Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, *36*, 1–17.

Salazar, A., Wuhrer, S., Shu, C., & Prieto, F. (2014). Fully automatic expression-invariant face correspondence. *Machine Vision and Applications*, *25*, 859–879.

Savran, A., Alyüz, N., Dibeklioğlu, H., Çeliktutan, O., Gökberk, B., Sankur, B., & Akarun, L. (2008). Bosphorus database for 3d face analysis. In B. Schouten, N. C. Juul, A. Drygajlo, & M. Tistarelli (Eds.), *Biometrics and identity management* (pp. 47–56). Springer.

Shimada, S., Golyanik, V., Tretschk, E., Stricker, D., & Theobalt, C. (2019). DispVoxNets: Non-rigid point set alignment with supervised learning proxies. In *Proceedings—2019 international conference on 3D vision, 3DV 2019*. https://doi.org/10.1109/3DV.2019.00013.

Szegedy, C., Wei, L., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *2015 IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE (pp. 1–9).

Tam, G. K., Cheng, Z. Q., Lai, Y. K., Langbein, F. C., Liu, Y., Marshall, D., et al. (2013). Registration of 3d point clouds and meshes: A survey from rigid to nonrigid. *IEEE Transactions on Visualization and Computer Graphics*, *19*(7), 1199–1217.

Tena, J. R., De La Torre, F., & Matthews, I. (2011). Interactive region-based linear 3D face models. *ACM Transactions on Graphics*, *30*(4), 1–10.

Tran, L., & Liu, X. (2018). Nonlinear 3D face morphable model. In *2018 IEEE/CVF conference on computer vision and pattern recognition*. IEEE (pp. 7346–7355).

Tran, L., Liu, F., & Liu, X. (2019). Towards high-fidelity nonlinear 3D face morphable model. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lió, P., & Bengio, Y. (2018). Graph attention networks. In *ICLR*.

Verma, N., Boyer, E., & Verbeek, J. (2018). FeaStNet: Feature-steered graph convolutions for 3D shape analysis. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*.

Vlasic, D., Brand, M., Pfister, H., & Popović, J. (2005). Face transfer with multilinear models. *ACM Transactions on Graphics*, *24*(3), 426–433.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., & Liu, W. (2018). CosFace: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*.

Wang, Y., & Solomon, J. (2019a). Deep closest point: Learning representations for point cloud registration. In *Proceedings of the IEEE international conference on computer vision*.

Wang, Y., & Solomon, J. M. (2019b). PRNet: Self-supervised learning for partial-to-partial registration. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32, pp. 8814–8826). Curran Associates Inc.

Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., & Solomon, J. M. (2019). Dynamic graph CNN for learning on point clouds. *ACM Transactions on Graphics*, *38*(5), 146:1–146:12.

Wu, Y., & He, K. (2020). Group normalization. *International Journal of Computer Vision*, *128*(3), 742–755.

Xu, Y., Fan, T., Xu, M., Zeng, L., & Qiao, Y. (2018). SpiderCNN: Deep learning on point sets with parameterized convolutional filters. In *The European conference on computer vision (ECCV) 11212 LNCS* (pp. 90–105).

Yin, L., Wei, X., Sun, Y., Wang, J., & Rosato, M. J. M. (2006). A 3D facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*. IEEE (pp. 211–216).

Yin, L., Chen, X., Sun, Y., Worm, T., & Reale, M. (2008). A high-resolution 3D dynamic facial expression database. In *2008 8th IEEE international conference on automatic face and gesture recognition, FG 2008*. IEEE (pp. 1–6).

Zhang, Z., Hua, B. S., & Yeung, S. K. (2019). Shellnet: Efficient point cloud convolutional neural networks using concentric shells statistics. In *International conference on computer vision (ICCV)*.

Zhu, X., Lei, Z., Yan, J., Yi, D., & Li, S. Z. (2015). High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.