PointAE: Point Auto-encoder for 3D Statistical Shape and Texture Modelling

Hang Dai, Ling Shao
Inception Institute of Artificial Intelligence
Abu Dhabi, UAE
hang.dai@inceptioniai.org, ling.shao@ieee.org

Abstract

The outcome of standard statistical shape modelling is a vector space representation of objects. Any convex combination of vectors of a set of object class examples generates a real and valid example. In this paper, we propose a Point Auto-Encoder (PointAE) with skip-connection, attention block for 3D statistical shape modelling directly on 3D points. The proposed PointAE is able to refine the correspondence with a correspondence refinement block. The data with refined correspondence can be fed to the PointAE again and bootstrap the constructed statistical models. Instead of two separate models, PointAE can simultaneously model the shape and texture variation. The extensive evaluation in three open-sourced datasets demonstrates that the proposed method achieves better performance in representation ability of the shape variations.

1. Introduction

With prior knowledge and experience, people can easily observe rich shape and texture variation for a certain type of object, such as human faces, cats or chairs, in both 2D and 3D images. This ability helps us recognise the same person, distinguish different kinds of creatures and sketch unseen samples of the same object class. The process of capturing this prior knowledge is mathematically interpreted as statistical modelling. One such outcome is a 3D morphable model (3DMM), a vector space representation of objects, that captures the variation of shape and texture. Any convex combination of vectors of a set of object class examples generates a real and valid example in this vector space. Statistical shape modelling is extensively studied in a variety of disciplines; for example computer vision, where researchers focus on applications in medical imaging, biometrics and the creative industries.

Statistical shape modelling aims to characterise the mean shape, and the variances and covariances of different object parts for various classes of object. Shape, as defined by D.G. Kendall [26], is all the geometrical information that remains when location, scale and rotational effects are filtered out from an object. In other words, those similarity effects need to be filtered out by aligning a collection of shape when doing shape analysis. A shape is described by locating a number of points on the outline. These points are defined as points of correspondence on each object that matches between and within populations. Statistical shape modelling is perhaps most commonly performed by Principal Component Analysis (PCA) over a set of meshes, which finds the directions in the vector space that have maximum variance, whilst being mutually orthogonal. However, PCA filters out high frequency signals, thereby losing shape detail in shape reconstruction. To overcome this, we employ point auto-encoder to extract latent representation of shapes and reconstruct the shapes.

The latent representation is a distributed representation that captures the coordinates along the main factors of variation in the data. This is similar to the way the projection on principal components would capture the main factors of variation in the data. Indeed, if there is one linear hidden layer and the mean squared error criterion is used to train the network, then the m hidden units learn to project the input in the span of the first m principal components of the data. If the hidden layer is non-linear, the auto-encoder behaves differently from PCA, with the ability to capture multi-modal aspects of the input distribution.

We propose a deep method to model 3D shape. In particular, our approach features a novel Point Auto-Encoder (PointAE) with skip-connection and attention block with contributions in:

- Unlike the previous deep methods which requires transferring the mesh into other geometric representations/features or remeshing the surface, 3D points can be directly fed to the proposed pointAE for statistical shape modelling.
- We propose a correspondence refinement block following PointAE to refine the correspondence. Then we bootstrap the modelling process by feeding the data with refined correspondence to PointAE. The ablation study shows that this process enhanced the shape rep-

resentation ability.

- Instead of modelling shape and texture separately, PointAE can treat an input which includes 3D coordinates XYZ and texture color RGB as a point cloud with 6 channels. To our best knowledge, our PointAE is the only deep method that models shape and texture simultaneously.
- We apply the proposed PointAE to three types of public datasets, which are face, head and body dataset, respectively. Both qualitative and quantitative evaluation demonstrate that the proposed method improves performance over the current state-of-art methods.

2. Related Work

2.1. History of 3DMM

In the 1990s, Cootes et al. developed shape models applied to 2D images, termed Point Distribution Models (PDMs) [11]. The work is done with reference to 2D shapes, where corresponding points are manually marked on the boundaries of a set of training examples. Cootes et al. presented Active Shape Models (ASM) in [12], where pose, scale and shape parameters are determined in order to fit the model to an image. This work was inspired by the earlier work on active contour models [25]. The same research team also went on to include texture in their models to give active appearance models [10]. They developed a set of shape modelling approaches where the best correspondences are those that define the most compact shape model given some quality of fit between the model and the data [16, 28]. Terzopoulos and Metaxas [39] introduced a physically-based approach to fitting 3D shapes. They formulated deformable superquadrics which incorporate the global shape parameters of a conventional superellipsoid with the local degrees of freedom of a spline. Kakadiaris et al. [24] presented an integrated approach to do shape segmentation and motion estimation using a physics-based framework.

Existing 3D statistical face models mainly consist of either morphable models, multilinear models and part-based models. In the late 1990s, Blanz and Vetter built a 3DMM from 3D face scans [3] and employed it in 2D face recognition [4]. Two hundred scans were used to build the model (young adults, 100 males and 100 females). The Basel Face Model (BFM) is the most well-known and widely-used and was developed by Paysan et al. [31]. The part-based model was shown to lead to a higher data accuracy than the global model [38, 2].

A statistical model called the multi-linear model [42, 44, 5, 43] is employed to statistically model the varying facial expressions. By using a multi-linear model, Vlasic et al. [42] modelled facial shape using a combination of identity and expression variation. Yang et al. [44] modelled the expression of a face in a different input image of the same subject. A number of PCA shape spaces for each expression are built and combined with a multi-linear model. A follow-up work [5, 43] used this model for a better description of expressions in videos.

A hierarchical pyramids method was introduced by Golovinskiy et al. to build a localised model [20]. In order to model the geometric details in a high resolution face mesh, this statistical model is able to describe the varying geometric facial detail. Brunton et al. [7] described 3D facial shape variation at multiple scales using wavelet basis. The wavelet basis provided a way to combine small signals in local facial regions which are difficult for PCA to capture. Claes et al. [9] explored the independent effects of the sex, genomic ancestry and genotype on facial shape variation. The experimental results showed that a set of 20 genes has significant effects on facial shape variation.

In 2017, Booth et al. [6] built a Large Scale Facial Model (LSFM), using the nonrigid iterative closest point template morphing approach, as was used in the BFM, but with error pruning, followed by Generalised Procrustes Analysis (GPA) for alignment, and PCA for the model construction. This 3DMM employs the largest 3D face dataset to date, and is constructed from 9663 distinct facial identities. Marcel et al. [30] model the shape variations with a Gaussian process, which they represent using the leading components of its Karhunen-Loeve expansion. This Gaussian Process Morphable Models (GPMMs) unify a variety of non-rigid deformation models with B-splines and PCA models as examples. In their follow-on work, they present a novel pipeline for morphable face model construction based on Gaussian processes [19]. GPMMs separate problem-specific requirements from the registration algorithm by incorporating domain-specific adaptions as a prior model.

Tran et al. [41] proposed a framework to construct a nonlinear 3DMM model from a large set of unconstrained face images, without collecting 3D face scans. Specifically, given a face image as input, a network encoder estimates the projection, shape and texture parameters. Two decoders served as the nonlinear 3DMM to map from the shape and texture parameters to the 3D shape and texture, respectively.

2.2. Deep 3D Shape Modelling

Genova et al. [18] presented a method for training a regression network from image pixels to 3D morphable model coordinates, where supervised training data is not necessary. Tewari et al. [40] fused a convolutional encoder with a differentiable renderer and a self-supervised training loss in a end-to-end training framework. Kim et al. [27] employed a deep convolutional inverse rendering framework for faces that aimed at estimating facial pose, shape, expression, reflectance and illumination, by estimating all parameters from just a single image.

Recently Bagautdinov et al. [1] proposed a method to model multi-scale face geometry that learns the facial geometry using UV parameterization for mesh representation. Tan et al. [37] employed mesh variational auto-encoders to explore the probabilistic latent space of 3D meshes. The training is performed on surface representation called RIMD (Rotation Invariant Mesh Difference) rather than the UV parameterization for the mesh. Ranjan et al. [34] introduced a Convolutional Mesh Autoencoder (CoMa) consisting of mesh downsampling and mesh upsampling layers with fast localised convolutional filters [17] defined on the mesh surface. This requires remeshing the surface and considering the triangulation relation. It is more difficult to model the dynamic high resolution 3D meshes [8, 29] when taking the temporal correlation into consideration.

The recent progress in 3DMM construction has a trend in applying 3D deep learning to model the nonlinear shape variations. Such variations are usually unable to be statistically described by traditional modelling methods. The previous methods described above need to remesh the surface, consider connectivity relation or transfer mesh into other geometric representations. In contrast, we propose a novel point auto-encoder based on PointNet [33] architecture to directly consume 3D point for statistical shape modelling. Moreover, these deep methods stick to only shape modelling.

3. Method

We propose a point auto-encoder for 3D statistical shape and texture modelling. In the following sections, we first describe the data preprocessing which is used for preparing the input data for the PointAE. We then formulate the methodology of PointAE mathematically. Following this we present the architecture of the proposed PointAE, which is followed by a correspondence refinement block. The following section is used for the description of simultaneous shape and texture modelling using PointAE. Finally, we present the implementation detail of our method.

3.1. Data Preprocessing

The statistical modelling process is feasible if and only if each mesh is reparametrised into a consistent form where the number of vertices, the triangulation, and the (approximate) anatomical meaning of each vertex are made consistent across all meshes. For example, given a vertex with index i in one mesh corresponding to the left mouth corner, it is required that the vertex with the same index in every mesh should correspond to the left mouth corner too. Meshes, every vertex of which satisfies the above properties, are said to be in dense correspondence with one another. We use the template morphing method from [13] to build the dense correspondence in Headspace dataset and the template morphing method from [19] in BU3DFE dataset. Caesar dataset provides registered scans with dense correspondence.

Once dense correspondence is established, the registered data shares the same triangulation relationship across the dataset. So we can take this triangulation relationship out of the statistical modelling process and only use 3D points for modelling. The collection of scans in dense correspondence are then subjected to Generalised Procrustes Analysis (GPA) [21] to remove similarity effects (rotation, and translation), leaving only shape information for modelling.

3.2. Problem Formulation

A global linear model such as the one of [23] represents all possible face shapes as linear combinations in a set of basis vectors. In [14, 15], it was obtained by performing principal component analysis on a training database.

In this paper, the proposed PointAE consumes point cloud directly to decompose the dataset into a latent representation. The dense correspondence is already established for each 3D scan. Each densely aligned 3D scan $X \in \mathbb{R}^n$ has n points:

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_i, \cdots, \mathbf{x}_{n-1}, \mathbf{x}_n], \tag{1}$$

where $\mathbf{x}_i$ indicates the i-th point, the value of which is the 3D coordinates $\mathbf{x}_i = (x_i, y_i, z_i)$. Formally, the point auto-encoder can be formulated as:

$$\mathbf{L} = E(\mathbf{X}; \theta_e), \quad \mathbf{X}^* = D(\mathbf{L}; \theta_d) \tag{2}$$

where $E(.; \theta_e)$ and $D(.; \theta_d)$ are multi-layer convolutional encoder and decoder, parameterised by weights θ_e and θ_d respectively, $\mathbf{L} \in \mathbb{R}^k$ is a set of k latent parameters, and $\mathbf{X}^* \in \mathbb{R}^n$ is the point cloud reconstructed from decoder, such that $||\mathbf{X} - \mathbf{X}^*||$ is minimised.

3.3. Point Auto-encoder

We propose a Point Auto-encoder (PointAE) to perform statistical shape modelling directly on 3D points. As shown in Figure 1, the input point set $\in \mathbb{R}^{n \times 3}$ is encoded into a latent representation which is then decoded towards 3D coordinates. The main block of encoder consists of a convolution layer followed by batch normalisation, concatenat layer with skip connection and a attention block based on [22]. At end of the four main blocks, the output is max-pooled to a flat vector, which is a compact representation of the input 3D points. This is also called the latent variables/representations of the input data. The decoder uses three fully connected (FC) layers to upsample the latent representation towards $3n$ dimensional features, which is then reshaped as the output point set $\in \mathbb{R}^{n \times 3}$.

Loss Function. Since the correspondence is established before training, we can calculate the mean per-point error

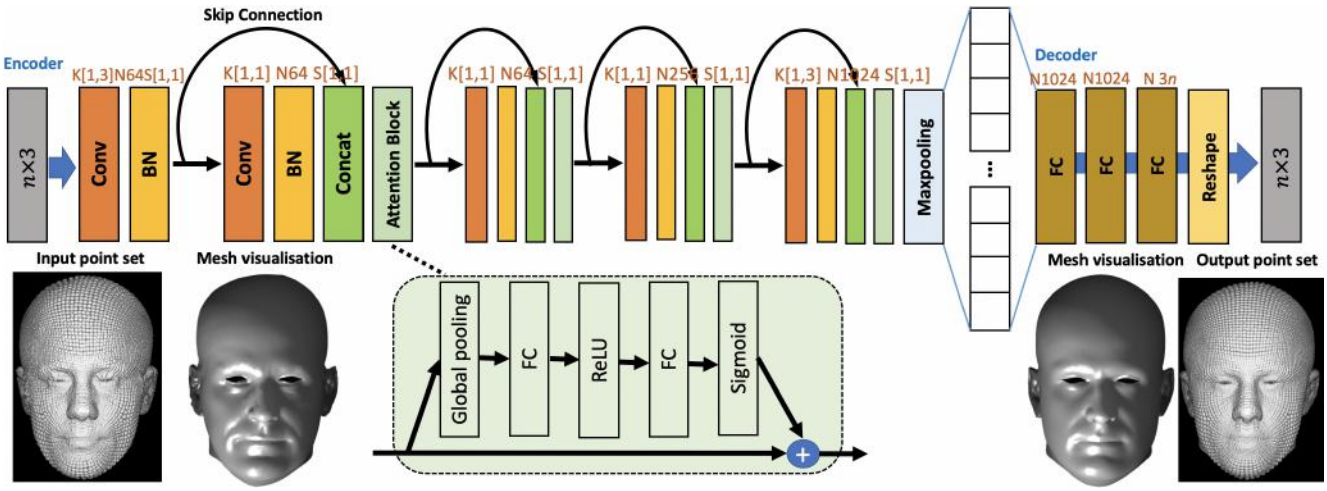

Figure 1. The Architecture of PointAE Network for 3D statistical modelling with corresponding kernel size (K), number of feature maps (N) and stride (S) indicated for each convolutional layer.

when minimising $||\mathbf{X} - \mathbf{X}^*||$:

$$\ell(\mathbf{X}^*, \mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\sum_{\mathbf{x}_i \in \mathbf{X}, \mathbf{x}_i^* \in \mathbf{X}^*}^{xyz} (\mathbf{x}_i - \mathbf{x}_i^*)^2} \qquad (3)$$

This is the key task in the proposed PointAE that is to minimise the distance error between $\mathbf{x}_i$ and $\mathbf{x}_i^*$.

3.4. Correspondence Refinement Block

The auto-encoder is able to denoise the data. We exploit this property to refine the correspondence and bootstrap the statistical modelling. As shown in Figure 2, the proposed PointAE refine the mesh structure of the failure cases. Because the latent representation $\mathbf{L}$ is viewed as a lossy compression of $\mathbf{X}$, it can not be a good (small-loss) compression for arbitrary inputs. That is the sense in which an auto-encoder generalizes: it gives low reconstruction error on test examples from the same distribution as the training examples, but it generates the best reconstruction for the arbitrary parts following this distribution. This mechanism enables the auto-encoder to refine the correspondence. The proposed PointAE forces points to follow the same distribution regularity and refine the noisy points using the knowledge from the training examples. The proposed correspondence refinement block is expected to achieve two goals together: (1) keep the same distribution regularity; (2) decrease distance error between the observation $\mathbf{X}$ and the reconstruction $\mathbf{X}^*$.

To do this, we use Laplace-Beltrami (LB) regularised mesh manipulation to retain the mesh structure when moving towards the 3D raw scan. Given the vertices of a scan stored in the matrix $\mathbf{X}_{\text{input}} \in \mathbb{R}^{n\times 3}$ and the denoised mesh from PointAE whose vertices are stored in the matrix $\mathbf{X}_{\text{denoised}} \in \mathbb{R}^{n\times 3}$, we define the selection matrices $\mathbf{S}_1 \in [0,1]^{m\times n}$ and $\mathbf{S}_2 \in [0,1]^{m\times n}$ as those that select the m vertices with mutual nearest neighbours from denoised mesh and the input scan respectively. This correspondence refinement system can be written as:

$$\begin{pmatrix} \lambda\mathbf{C} \\ \mathbf{S}_1 \end{pmatrix} \mathbf{X}_{\text{refined}} = \begin{pmatrix} \lambda\mathbf{C}\mathbf{X}_{\text{denoised}} \\ \mathbf{S}_2\mathbf{X}_{\text{input}} \end{pmatrix} \qquad (4)$$

where $\mathbf{C} \in \mathbb{R}^{p\times p}$ is the cotangent Laplacian approximation to the LB operator [36] and $\mathbf{X}_{\text{refined}} \in \mathbb{R}^{p\times 3}$ are the refined vertex positions that we wish to solve for. The parameter λ weights the relative influence of the position and regularisation constraints, effectively determining the 'stiffness' of the projection. As $\lambda \to 0$, the projection tends towards nearest neighbour projection. As $\lambda \to \infty$, the deformed template will only be allowed to rigidly transform. After correspondence refinement, the point sets can be fed to PointAE. As shown in Figure 2, the points with better correspondence from the bootstrapping process are fed to PointAE.

3.5. Shape and Texture Modelling

Unlike previous methods using one other model for texture modelling, the proposed method is able to model shape and texture simultaneously while not increasing the number of latent parameters. To do this, we formulate n dimensional input with 6 channels as:

$$\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_i, \cdots, \mathbf{q}_{n-1}, \mathbf{q}_n], \qquad (5)$$

where the input $\mathbf{Q} \in \mathbb{R}^n$ and the output $\mathbf{Q}^* \in \mathbb{R}^n$ have n points with 6 channels, $\mathbf{q}_i = (x_i, y_i, z_i, r_i, g_i, b_i)$ represents the combination of 3D coordinates XYZ and texture color RGB. So the proposed method can treat this input as a 6-channel point set, which is normalised between $[0, 1]$. Then the point auto-encoder can be formulated as:

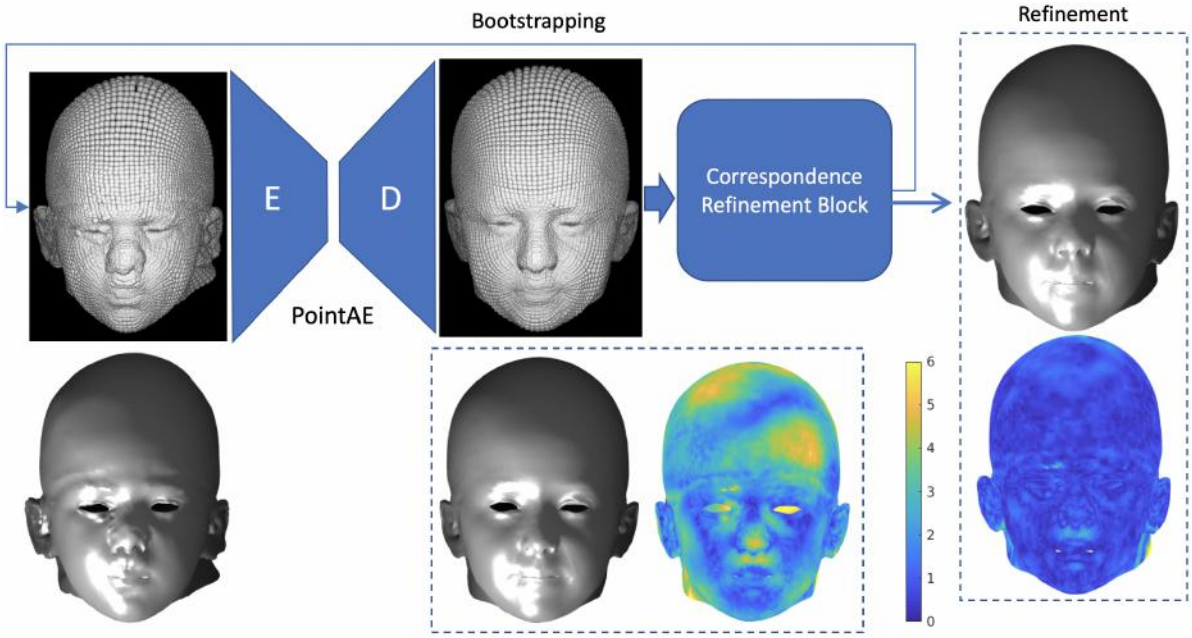

Figure 2. Correspondence refinement block. Note that PointAE refines the mesh structure in nose and mouth region, but it has noticeable distance error in nose, mouth and cranium region. The correspondence refinement block significantly decreases the distance error and retain mesh structure as rigid as possible as the $\mathbf{X}^*$.

$$\mathbf{L} = E(\mathbf{Q}; \theta_e), \quad \mathbf{Q}^* = D(\mathbf{L}; \theta_d) \tag{6}$$

When $||\mathbf{Q} - \mathbf{Q}^*||$ is minimised, the encoder generates the latent variables to compactly represent both shape and texture. However, the standard method–PCA needs to flat the input as $6n$ dimensional input.

Loss Function. For the 3D coordinates xyz, we can calculate the mean per-point error, while we employ root mean square error (RMSE) for texture reconstruction loss. To minimise $||\mathbf{Q}-\mathbf{Q}^*||$, we can combine the mean per-point error and RMSE as the loss function:

$$\ell^{st}(\mathbf{Q}^*, \mathbf{Q}) = \ell(\mathbf{Q}^*_{xyz}, \mathbf{Q}_{xyz}) + RMSE(\mathbf{Q}^*_{rgb}, \mathbf{Q}_{rgb}) \tag{7}$$

So the proposed PointAE can statistically model the 3D shape and texture simultaneously.

3.6. Implementation Details

PointAE. The Architecture of PointAE network is shown in Figure 1 with corresponding kernel size (K), number of feature maps (N) and stride (S) indicated for each convolutional layer. For the encoder, the structure is as follows: (1) K[1,3] N64 S[1,1], (2) K[1,1] N64 S[1,1], (3) K[1,1] N64 S[1,1], (4) K[1,1] N256 S[1,1] and (5) K[1,1] N1024 S[1,1]. The 1024 features are concatenated with 256 features from the skip connection. The merged 1280 features are maxpooled to a latent space, which has 1280 latent variables/representations. For the FC decoder, the structure is as follows: (1) N1024, (2) N1024 and (3) N 3n. The final layer reshape the $3n$ dimensional features into $n \times 3$ points, which is the reconstruction.

Training. We trained the model in DGX1 machine using the following settings: batch size is 64, 2001 epochs, and the learning rate starts at 0.001. We used the adam optimiser and $\lambda = 1$.

4. Experimental Results

We evaluate the proposed method in three open datasets: BU3DFE, Headspace and Caesar, which are face, full head and body dataset. We compare the proposed method with both standard method and deep methods. Note that among the deep methods for comparison, the proposed method is the only one which consume 3D point directly without transferring mesh into other geometric representations/features or remeshing the surface. We evaluate the proposed method qualitatively and quantitatively.

4.1. Evaluation Metric

The accuracy is evaluated by the Normalized Mean Error (NME), that is the average of per-vertex distance error normalized by the size of the 3D mesh. After the mesh is pose normalised, the size can be defined as the maximum value of length (l), width (w) and height (h) as demonstrated in Figure 3. So the NME can be formulated as:

$$NME(\mathbf{X}^*, \mathbf{X}) = \frac{1}{n \times max(l, w, h)} \ell(\mathbf{X}^*, \mathbf{X}) \tag{8}$$

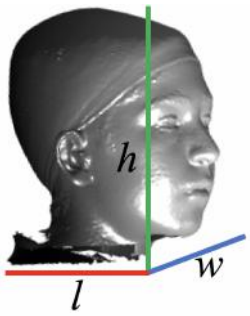

Figure 3. Demonstration of length (l), width (w) and height (h) in 3D mesh.

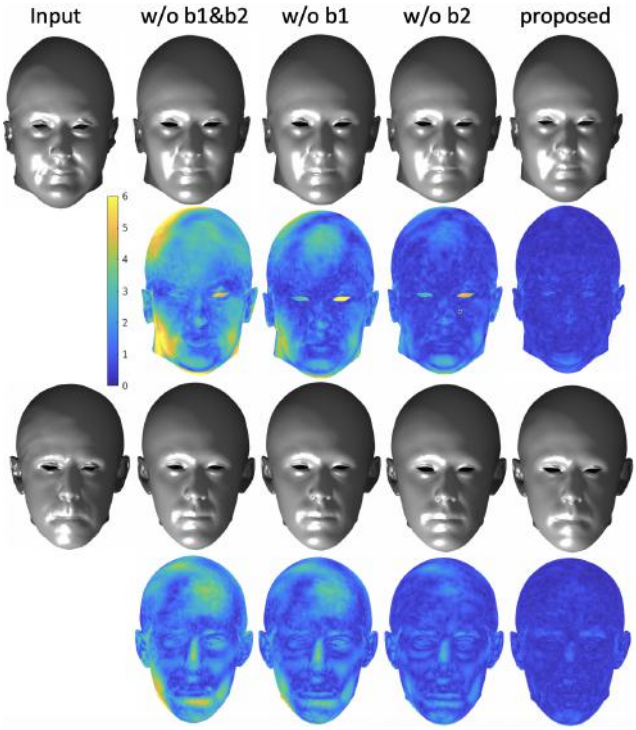

Figure 4. Reconstruction results with error map: (1) input data; (2)w/o b1&b2; (3) w/o b1; (4) w/o b2 and (5) proposed method.

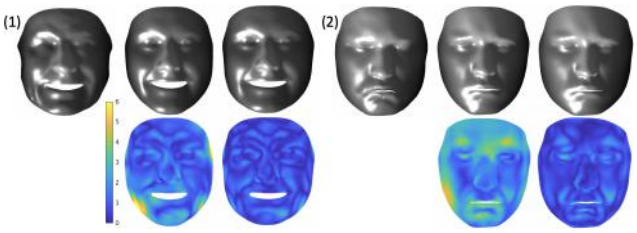

Figure 5. Reconstruction results with error map on BU3DFE dataset: (1) and (2) input data, results from Compositional VAE and results from the proposed method.

4.2. Ablation Study

Following the setting in [14], we evaluate the function of each component such as skip connection (b1) and attention block (b2) in terms of reconstruction error. We use 80% of Headspace dataset for training and 20% for testing. We compute the normalised average per-vertex distance error for quantitative evaluation. Figure 4 demonstrates the reconstruction results along with error for the proposed method without (w/o) b1 and b2, without b1, without b2, the proposed method without correspondence refinement (p1) and the proposed method (p2). As shown in Table 1, the skip connection and attention block have significant improvement in the proposed method. In particular, the skip connection has more influence in the proposed method than the attention block. When comparing p1 and p2, p2 obtains less reconstruction error across all the number of the latent variables, which implies that the correspondence refinement block enhanced the shape representation ability.

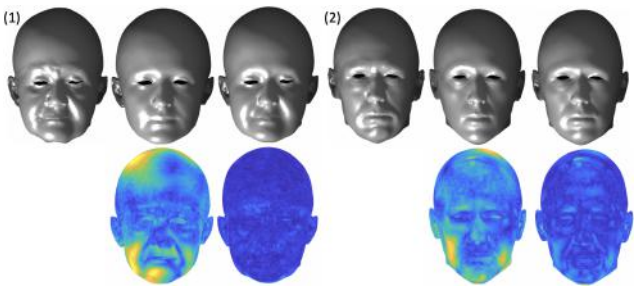

Figure 6. Reconstruction results with error map on Headspace dataset: (1) and (2) input data; reconstruction from PCA; (3)proposed method.

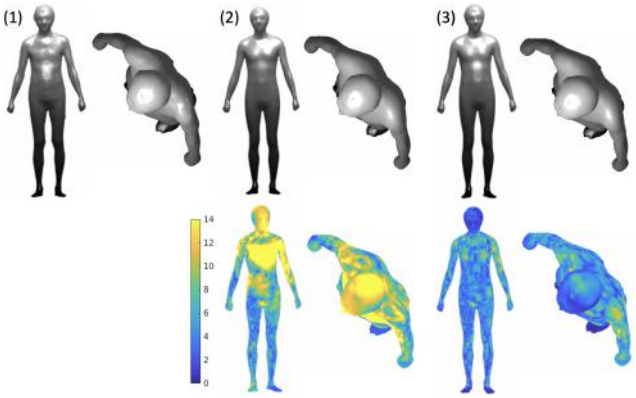

Figure 7. Reconstruction results with error map on Caesar dataset from frontal view and top view: (1) input data; (2) PCA; (3)proposed method.

# of L	w/o b	w/o b1	w/o b2	p1	p2
128	0.0612	0.0537	0.0271	0.0202	**0.0168**
256	0.0581	0.0517	0.0249	0.0163	**0.0143**
512	0.0540	0.0487	0.0225	0.0124	**0.0095**
1024	0.0501	0.0438	0.0196	0.0107	**0.0083**

Table 1. Normalized mean error on held-out shapes with different latent dimension.

Figure 8. Shape variations from the mean to mean + and - 3 standard deviations of top eight elements for the latent representations in Headspace dataset.

Figure 9. Shape interpolation to describe the ability of compressing expression: 1st row–PCA, 2nd row–proposed method.

4.3. Representation Power

To quantify the representation power of the statistical modelling methods, we calculate the normalised average per-vertex distance error on held-out shapes with different latent dimension. We compare the proposed method with PCA, MeshVAE [37] and Compositional VAE [1] in three types of open-sourced datasets: BU3DFE [45], Headspace [14, 35] and Caesar [32]. We use 80% of the dataset for training and 20% for testing.

Face. BU3DFE includes 100 subjects with 2500 facial expression models. Each subject contains one neutral and six expressions with four levels of strength. As can be seen from Figure 5, the reconstruction results with error map shows that for the two unseen examples, the proposed method has less reconstruction error, especially in chin, cheek and forehead region. Table 2 demonstrates the normalised average per-vertex distance error on held-out shapes with different latent dimension in BU3DFE. The proposed method achieves the best performance across all the number of latent variables. The reason why proposed method is better than Compositional VAE [1] is that PointAE is directly applied to 3D points and Compositional VAE [1] is to 2D UV representation of mesh.

Head. The Headspace dataset is a set of 3D images of the human head, consisting of 1519 subjects wearing tight fitting latex caps to reduce the effect of hairstyles. Figure 6 shows the reconstruction results with error map. For the two unseen shapes, the reconstruction of the proposed method is closer to the input data. In particular, the proposed method is able to capture the shape detail, for example, the wrinkles around mouth. Table 3 demonstrates the normalised average per-vertex distance error on held-out shapes with different latent dimension in Headspace. The proposed method obtains a slight improvement than Compositional VAE [1]. More significant improvement exists when compared with PCA [14] in Headspace. This lies in the fact that nonlinear variations are captured in the five main blocks of decoder in PointAE, while the nonlinear variations are easy to be filtered out by PCA.

# of L	PCA [13]	[1]	Proposed
128	0.0311	0.0222	**0.0178**
256	0.0258	0.0174	**0.0147**
512	0.0213	0.0120	**0.0096**
1024	0.0187	0.0104	**0.0074**

Table 2. Normalized mean error on held-out shapes with different latent dimension in BU3DFE.

# of L	PCA [14]	[1]	proposed
128	0.0334	0.0194	**0.0168**
256	0.0286	0.0154	**0.0143**
512	0.0249	0.0124	**0.0095**
1024	0.0187	0.0103	**0.0083**

Table 3. Normalized mean error on held-out shapes with different latent dimension in Headspace.

Body. Caesar has 4309 full body registered scans with large variations in age distribution, weight and height, which makes it challenging to model the shape variation. In Figure 5, MeshVAE [37] shows a incorrect head pose and thinner chest, while the proposed method has less error across the whole body. As can be seen from Table 4, we compute the normalised average per-vertex distance error on held-out shapes with different latent dimension in Caesar. The proposed method obtains the lowest per-vertex reconstruction error on held-out shapes across all the choices

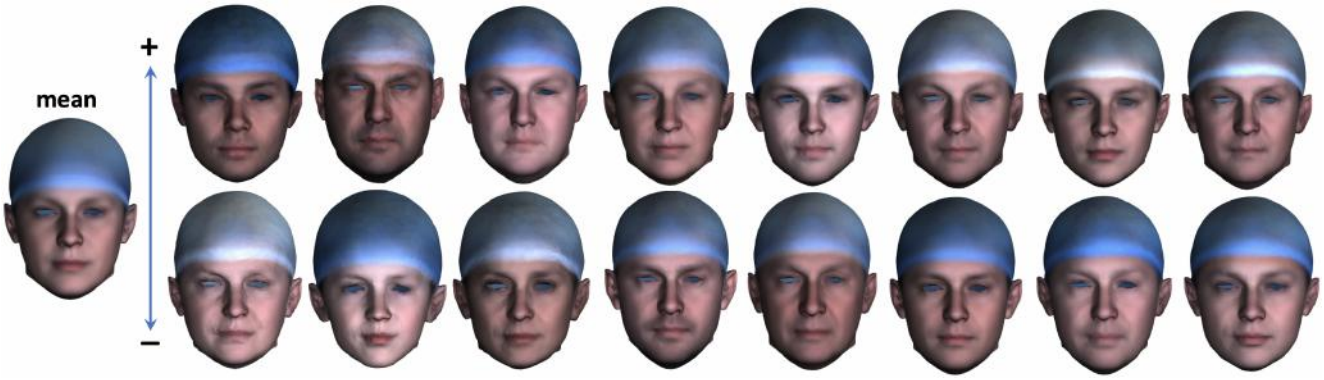

Figure 10. Texture and shape variations from the mean to mean + and - 3 standard deviations of top eight elements for the latent representations in Headspace dataset.

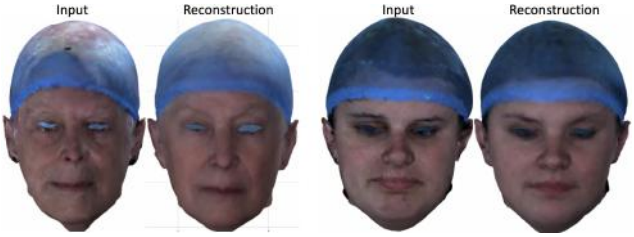

Figure 11. Two unseen examples reconstructed by shape&texture model: the re-scaled per-vertex reconstruction error for the first and second examples are 1.97 mm and 2.42 mm; The re-scaled ([0 255]) texture RMSE is 2.68 and 3.12.

# of L	PCA [32]	MeshVAE [37]	proposed
128	0.0735	0.0548	**0.0503**
256	0.0648	0.0511	**0.0459**
512	0.0581	0.0448	**0.0409**
1024	0.0513	0.0405	**0.0378**

Table 4. Normalized mean error on held-out shapes with different latent dimension in Caesar.

of the number of latent variables. This has larger error than face and head modelling because of body size.

4.4. Morphable Models

We can build the shape model and texture model simultaneously with the proposed method. After reconstruction, we need to re-scale the 3D coordinates back from [0 1]. Since Headspace released both 3D shape and texture data, we use the proposed method to model both shape and texture variation in this dataset.

Shape model. We explore the learned latent representations by visualising the top eight components. We first encode all training samples (both shape and texture) in the latent space via the trained encoder. We compute the mean and the mean $+$ and $-$ 3 standard deviations of top eight components of the latent representation. We then perturb each element of the latent vector with the amount of perturbation equal to the corresponding standard deviation, and use the decoder to transform the perturbed latent vector to a reconstructed sample. As can be seen from Figure 8, the first dominate latent variables show shape variation reflecting the gender, while the second one demonstrates age correlated shape variation. With less dominate latent variables, the shape variations become smaller and smaller. We use shape interpolation from one expression to another for describing the ability of compressing expression. In Figure 9, the proposed method presents a more smooth shape interpolation than PCA model.

Shape&Texture model. We select the top eight elements for the latent representations. From the joint learning of shape and texture, we can view both the texture and shape variations together. the shape model the shape for visualisation of the texture model. As shown in Figure 10, it shows texture and shape variations from the mean to mean + and - 3 standard deviations of top eight elements for the latent representations. The first dominant texture variation is mainly from white to dark and the second is from young to old which has some moustache. Figure 11 demonstrates two unseen examples reconstructed by shape&texture model. The re-scaled per-vertex reconstruction error for the first and second examples are 1.97 mm and 2.42 mm. The re-scaled ([0 255]) texture RMSE is 2.68 and 3.12.

5. Conclusion

We proposed a PointAE to simultaneously perform 3D shape and texture statistical modelling directly on 3D points and texture. The proposed PointAE achieved lower reconstruction error compared with the state-of-art methods. The 3D models constructed by the proposed method are more powerful in shape representation ability. The proposed PointAE has the ability to refine the correspondence with the proposed correspondence refinement block. The one shot bootstrapping enhanced the representation ability of the constructed 3D models.

References

[1] Timur Bagautdinov, Chenglei Wu, Jason Saragih, Pascal Fua, and Yaser Sheikh. Modeling facial geometry using compositional vaes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3877–3886, 2018. 3, 7

[2] Curzio Basso, Alessandro Verri, and Jens Herder. Fitting 3d morphable models using implicit representations. *Journal of Virtual Reality and Broadcasting*, 4(18):1–10, 2007. 2

[3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 2

[4] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on pattern analysis and machine intelligence*, 25(9):1063–1074, 2003. 2

[5] Timo Bolkart and Stefanie Wuhrer. Statistical analysis of 3d faces in motion. In *2013 International Conference on 3D Vision-3DV 2013*, pages 103–110. IEEE, 2013. 2

[6] James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. A 3d morphable model learnt from 10,000 faces. In *Proceedings of CVPR*, pages 5543–5552, 2016. 2

[7] Alan Brunton, Jochen Lang, Eric Dubois, and Chang Shu. Wavelet model-based stereo for fast, robust face reconstruction. In *2011 Canadian Conference on Computer and Robot Vision (CRV)*, pages 347–354, 2011. 2

[8] Shiyang Cheng, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. 4dfab: A large scale 4d database for facial expression analysis and biometric applications. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5117–5126, 2018. 3

[9] Peter Claes, Denise K Liberton, Katleen Daniels, Kerri Matthes Rosana, Ellen E Quillen, Laurel N Pearson, Brian McEvoy, Marc Bauchet, Arslan A Zaidi, Wei Yao, et al. Modeling 3d facial shape from dna. *PLoS genetics*, 10(3):e1004224, 2014. 2

[10] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):681–685, 2001. 2

[11] Timothy F Cootes and Christopher J Taylor. Combining point distribution models with shape models based on finite element analysis. *Image and Vision Computing*, 13(5):403–409, 1995. 2

[12] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995. 2

[13] Hang Dai, Nick Pears, and William Smith. Non-rigid 3d shape registration using an adaptive template. In *European Conference on Computer Vision Workshop*, pages 48–63. Springer, 2018. 3, 7

[14] Hang Dai, Nick Pears, William Smith, and Christian Duncan. A 3d morphable model of craniofacial shape and texture variation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3104–3112. IEEE, 2017. 3, 6, 7

[15] Hang Dai, WA Smith, Nick Pears, and Christian Duncan. Symmetry-factored statistical modelling of craniofacial shape. In *Proceedings of the International Conference on Computer Vision Workshop Venice, Italy*, pages 22–29, 2017. 3

[16] Rhodri H Davies, Carole J Twining, Tim F Cootes, John C Waterton, and Camillo J Taylor. A minimum description length approach to statistical shape modeling. *Medical Imaging, IEEE Transactions on*, 21(5):525–537, 2002. 2

[17] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, pages 3844–3852, 2016. 3

[18] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8377–8386, 2018. 2

[19] Thomas Gerig, Andreas Forster, Clemens Blumer, Bernhard Egger, Marcel Lüthi, Sandro Schönborn, and Thomas Vetter. Morphable face models - an open framework. *CoRR*, abs/1709.08398, 2017. 2, 3

[20] Aleksey Golovinskiy, Wojciech Matusik, Hanspeter Pfister, Szymon Rusinkiewicz, and Thomas Funkhouser. A statistical model for synthesis of detailed facial geometry. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 1025–1034, 2006. 2

[21] Colin Goodall. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 285–339, 1991. 3

[22] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3

[23] IEEE. *A 3D Face Model for Pose and Illumination Invariant Face Recognition*, Genova, Italy, 2009. 3

[24] Ioannis A Kakadiaris, Dimitri Metaxas, and Ruzena Bajcsy. Active part-decomposition, shape, and motion estimation of articulated objects: A physics-based approach. In *CVPR*, volume 94, pages 980–984, 1994. 2

[25] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988. 2

[26] David G Kendall. A survey of the statistical theory of shape. *Statistical Science*, pages 87–99, 1989. 1

[27] Hyeongwoo Kim, Michael Zollhöfer, Ayush Tewari, Justus Thies, Christian Richardt, and Christian Theobalt. Inversefacenet: Deep single-shot inverse face rendering from a single image. *arXiv preprint arXiv:1703.10956*, 2017. 2

[28] Aaron CW Kotcheff and Chris J Taylor. Automatic construction of eigenshape models by direct optimization. *Medical Image Analysis*, 2(4):303–314, 1998. 2

[29] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics (TOG)*, 36(6):194, 2017. 3

[30] Marcel Lüthi, Thomas Gerig, Christoph Jud, and Thomas Vetter. Gaussian process morphable models. *IEEE transactions on pattern analysis and machine intelligence*, 2017. 2

[31] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *Advanced video and signal based surveillance, 2009. AVSS'09. Sixth IEEE International Conference on*, pages 296–301, 2009. 2

[32] Leonid Pishchulin, Stefanie Wuhrer, Thomas Helten, Christian Theobalt, and Bernt Schiele. Building statistical shape spaces for 3d human modeling. *Pattern Recognition*, 67:276–286, 2017. 7, 8

[33] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017. 3

[34] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3d faces using convolutional mesh autoencoders. *arXiv preprint arXiv:1807.10267*, 2018. 3

[35] B Robertson, H Dai, N Pears, and C Duncan. A morphable model of the human head validating the outcomes of an age-dependent scaphocephaly correction. *International Journal of Oral and Maxillofacial Surgery*, 46:68, 2017. 7

[36] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Proceedings of the Fifth Eurographics Symposium on Geometry Processing*, pages 109–116, 2007. 4

[37] Qingyang Tan, Lin Gao, Yu-Kun Lai, and Shihong Xia. Variational autoencoders for deforming 3d mesh models. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3, 7, 8

[38] Frank B ter Haar and Remco C Veltkamp. 3d face model fitting for recognition. In *European Conference on Computer Vision*, pages 652–664, 2008. 2

[39] Demetri Terzopoulos and Dimitri Metaxas. Dynamic 3d models with local and global deformations: Deformable superquadrics. In *Computer Vision, 1990. Proceedings, Third International Conference on*, pages 606–615. IEEE, 1990. 2

[40] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. *arXiv preprint arXiv:1712.02859*, 2, 2017. 2

[41] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. *arXiv preprint arXiv:1804.03786*, 2018. 2

[42] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. Face transfer with multilinear models. In *ACM Transactions on Graphics (TOG)*, volume 24, pages 426–433, 2005. 2

[43] Fei Yang, Lubomir Bourdev, Eli Shechtman, Jue Wang, and Dimitris Metaxas. Facial expression editing in video using a temporally-smooth factorization. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 861–868. IEEE, 2012. 2

[44] Fei Yang, Jue Wang, Eli Shechtman, Lubomir Bourdev, and Dimitri Metaxas. Expression flow for 3d-aware face component transfer. In *ACM Transactions on Graphics (TOG)*, volume 30, page 60, 2011. 2

[45] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 211–216. IEEE, 2006. 7