Construction of Multi-resolution Multi-organ Shape Model Based on Stacked Autoencoder Neural Network

1st Zhonghua Chen
School of Biomedical Engineering, Faculty of Electronic Information and Electrical Engineering
Faculty of Information Technology
Dalian University of Technology
University of Jyväskylä
Dalian, China, 116024
Jyväskylä, Finland, 40014
zhonghua.x.chen@student.jyu.fi

2nd Hongkai Wang
School of Biomedical Engineering, Faculty of Electronic Information and Electrical Engineering
Dalian University of Technology
Dalian, China, 116024
wang.hongkai@dlut.edu.cn

3rd Fengyu Cong
School of Biomedical Engineering, Faculty of Electronic Information and Electrical Engineering
Faculty of Information Technology
School of Artificial Intelligence, Faculty of Electronic Information and Electrical Engineering
Key Laboratory of Integrated Circuit and Biomedical Electronic System
Dalian University of Technology
University of Jyväskylä
Dalian, China, 116024
Jyväskylä, Finland, 40014
cong@dlut.edu.cn

4th Lauri Kettunen
Faculty of Information Technology
University of Jyväskylä
Jyväskylä, Finland, 40014
lauri.y.o.kettunen@jyu.fi

***Abstract*— The construction of statistical shape models (SSMs) is an important method in the field of medical image segmentation. Most SSMs are constructed by using traditional modeling methods based on principal component analysis (PCA), which cannot fully present the true deformation ability of models. To solve the insufficient deformation ability of SSMs, we propose a stacked autoencoder (SAE) neural network to construct a multi-resolution multi-organ shape model based on mouse micro-CT images, which can express more linear and non-linear deformations than SSMs based on PCA. The main advantage of this method is that the SAE neural network is simple and flexible and it can learn more deformation modes from training data. We have quantitatively compared the modeling performance of this method with the constructed SSMs based on PCA in terms of model generalization and specificity.**

Index Terms*— *statistical shape model, principal component analysis, mouse micro-CT images, stacked autoencoder neural network, multi-resolution multi-organ shape model

I. Introduction

The construction of statistical shape models (SSMs) is an important segmentation method in the field of medical image segmentation. The applications of medical image segmentation methods focus on the segmentation of biological tissues and organs. Some applications consist of liver segmentation for volume measurement [1], breast tumor segmentation for diagnosis [2], bone localization and segmentation [3], and automatic heart segmentation for treatment [4], etc. In the past few decades, SSM methods have been employed to construct segmentations for single specific organs, especially for research purposes. Only a few other modeling methods focused on the segmentation of multiple organs. The most important approach in constructing SSMs is the principal component analysis (PCA) based on eigenvalue decomposition [5]. Given a set of medical images, such as computed tomography (CT) images, PCA approaches are used to calculate the deformation components and corresponding deformation coefficients. Thereafter the deformation components are added to the mean shape of these medical images to obtain a traditional standard SSM. Based on this rationale, various techniques have been proposed to improve the calculation of deformation components. Many other features of images are added to the PCA process to construct more representative models for images: they include different variants based on active shape models [6] and active appearance models [7] proposed by Cootes et al. Later in 2009 Tresadern et al. [8] combined an MRF-based local shape model with a PCA-based global shape model for modeling and locating deformable objects. Wilms et al. [9] expanded this work to construct a two-dimensional (2D) multi-resolution multi-organ model for the segmentation of 2D CT images.

Although PCA-based shape modeling methods have been used widely in medical image segmentation and many

National Natural Science Fund of China (No. 81971693, 81401475), Science and Technology Innovation Fund of Dalian City (2018J12GX042), Fundamental Research Funds for the Central Universities (DUT19JC01). China Scholarship Council (No. 201806060163)

different improved variants have been proposed in the last three decades, each modification can only be used in a specific application. Their performance may degrade dramatically when there are outliers in the data of the training models. Due to the limited deformation ability of linear SSMs, some non-linear PCA methods have been proposed to represent variations of bending and rotation in a more natural fashion [10, 11, 12]. Furthermore, when there are large amounts of three-dimensional (3D) training models, the computer's memory requirements increase exponentially, which leads to the failure of model construction [9, 13].

On the other hand, in recent years autoencoder-based shape modeling approaches have been proposed. They can learn latent non-linear deformations of shape models and represent more refined expressions in deformation models. In 2018, Litany et al. used a variational autoencoder to learn the latent space of objects with non-rigid deformations to reconstruct missing parts of objects [14]. At the same time, Ranjan et al. proposed to use convolutional mesh autoencoders learning non-linear variations for the construction of 3D faces [15].

A trained autoencoder network is simple for model construction and it only needs to save a few trained network parameters compared with PCA-based methods. In this paper, we have stacked autoencoder networks to build an SAE neural network that is used to construct a multi-resolution multi-organ shape model based on abdominal mouse micro-CT images. We have compared the models obtained from the SAE neural network with traditional global SSMs and local SSMs in terms of model generalization and specificity performance.

II. Materials and Methods

A. Mouse Micro-CT Data

In this study, we have collected micro-CT images of 98 mice from the Molecular Imaging Centre of the University of California, Los Angeles. For neural networks, sufficient training data should be provided to train the network. Practically, there are certain challenges in the collection of medical images, and 98 mouse images can be regarded as large sample data in the domain of medical image analysis. The detailed operating rules of mouse micro-CT imaging and the parameter settings of imaging equipment are described in our previous study [16].

After we have collected 98 mouse micro-CT images, we need to preprocess these images to obtain the final training data. The specific process is as follows: (1) We invited small animal imaging experts to manually segment 98 mouse micro-CT images to obtain labels of the liver, spleen, left kidney, and right kidney. (2) Next, these labels have been converted into multi-organ surface meshes with the moving cube algorithm [17]. (3) After this, we chose a multi-organ surface mesh from the meshes as a reference template and used the point cloud registration method [18, 19] to register it to the other 97 meshes. Finally, each mesh has the same number of points and each point on the mesh corresponds to the same anatomical position. (4) We performed the same downsampling on all points on every mesh to obtain point cloud data that can be trained by the network.

B. Methods

1) Description of autoencoder network

An autoencoder network is an unsupervised neural network that learns input data and reconstructs the input data to the greatest extent. Normally, the simplest autoencoder has three layers: an input layer, a hidden layer, and an output layer. The number of nodes in the output layer is the same as the input layer. Since the input data is lossy and relevant during compression and decompression, the autoencoder network can learn compression and decompression functions that contain specific features of the data and can be used for data dimensionality reduction and feature extraction. In practical applications, an autoencoder network must try to obtain the most important features that can represent the input data.

An autoencoder consists of an encoder ϕ, and a decoder Ψ, which can be defined as follows:

$$\begin{gathered} \phi: X \rightarrow F \\ \Psi: F \rightarrow X \\ \phi, \Psi = \operatorname{argmin}_{\phi,\Psi} \|X - (\Psi \circ \phi) X\|^2 \end{gathered} \tag{1}$$

where *argmin* represents obtaining minimum values of ϕ, Ψ simultaneously.

In the simplest case, given one hidden layer, the encoder stage of an autoencoder takes the input $x \epsilon R^d = X$ and maps it to $h \in R^p = F$:

$$h = \sigma(Wx + b), \tag{2}$$

where h is referred to as code, latent variables, or latent representation, σ is an element-wise activation function such as a sigmoid function or a rectified linear unit, W is a weight matrix, and b is a bias vector.

The decoder stage of the autoencoder maps h to the reconstruction x' of the same shape x:

$$x' = \sigma'(W'h + b'), \tag{3}$$

where σ' is also an element-wise activation function, W' is a weight matrix and b' is a bias vector.

Autoencoders are trained to minimize reconstruction errors which are often referred to as the "loss":

$$L(x, x') = \|x - x'\|^2 = \|x - \sigma'(W'(\sigma(Wx + b)) + b')\|^2 \tag{4}$$

The basic structure of an autoencoder network is shown in Fig. 1. The first column is the input layer, the middle column is a hidden layer and the last column is the output layer, +1 indicates an offset constraint condition added to the corresponding layer.

2) Multi-resolution shape modeling based on SAE neural network

A stacked autoencoder (SAE) neural network is composed of multiple networks connected in series. In this paper, the purpose of using the SAE neural network to train input mesh data is to extract high-dimensional features of the data layer

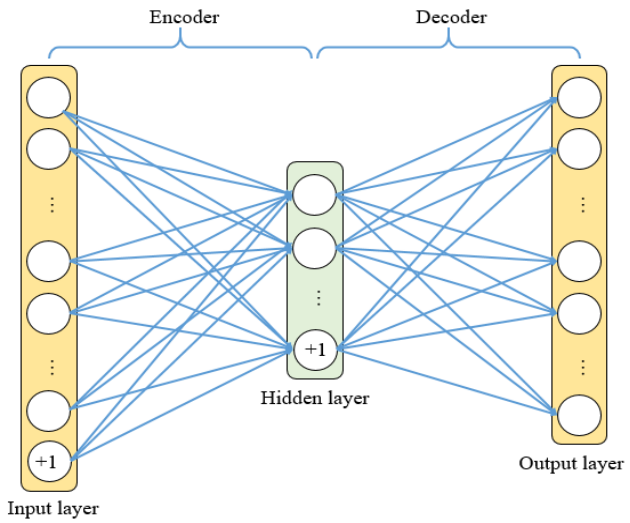

Fig. 1. Basic structure of an autoencoder network.

by layer and to reduce the dimension of the input data layer by layer. Therefore, the SAE neural network transforms high-dimensional mesh data into a series of low-dimensional feature vectors. Finally, these low-dimensional feature vectors are decompressed to reconstruct mesh data approximately equal to the input mesh data. In the process of decompressing features and restoring data, if we regularly modify high-dimensional feature parameters, we can see that the reconstructed data change regularly, which is also reflected in the deformation of 3D models.

Fig. 2 shows the multi-resolution shape modeling process based on SAE neural network. E represents a compression network (i.e., an encoder) and D represents a decompression network (i.e., a decoder). Each combination of E and D is an autoencoder network introduced in section 1).

This study uses an SAE neural network with three autoencoder networks. The training data for the first level E_1 and D_1 is the shape vector of meshes described in section A. After training, E_1 compresses the shape vector set into a shorter feature vector set $\vec{F}_1 = \{\vec{f}_{1,1}, \vec{f}_{1,2}, \dots, \vec{f}_{1,M}\}$, where M is the number of training meshes, and the subscript i, j of $\vec{f}_{i,j}$ represents the *j-th* sample of the *i-th* level. The $\vec{F}_1$ obtained from the first level network is used as the training sample of

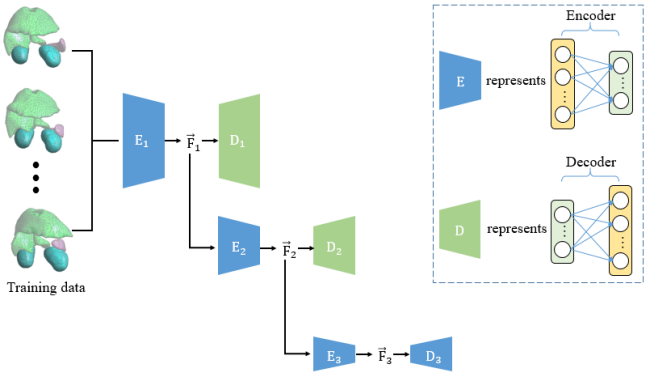

Fig. 2. The overall structure of a three-level SAE network with 3 autoencoder networks.

the second level network to obtain the training results of E_2 and D_2. The second level network generates a further compressed feature vector set $\vec{F}_2 = \{\vec{f}_{2,1}, \vec{f}_{2,2}, \dots, \vec{f}_{2,M}\}$. Then $\vec{F}_2$ is used to train the third level network (i.e., E_3 and D_3). As a result, the vector lengths of $\vec{F}_1$, $\vec{F}_2$, and $\vec{F}_3$ decrease gradually, which means the compression ratio gradually increases. From the perspective of shape characteristics, the resolution of the shape deformation should gradually transition from a global model to a local part of the model, thus forming a multi-resolution training method.

The SAE neural network used in this research encounters the problem of too large model data. Each 3D training mesh contains 3759 surface sampling points and the shape vector of a mesh contains 11277 dimensions, which takes up a lot of memory and causes a computational burden for SAE neural network training. Due to the limitation of the memory capacity of the computer used in this study, the points of each training mesh are downsampled to 752 as described in section A. Then the down-sampled model vertices are trained by the SAE neural network, and the generated deformation vectors are interpolated through the Laplace iteration diffusion algorithm (5) to obtain deformation vectors of 3759 vertices of a mesh.

$$\vec{X}_{i_{(n+1)}} = \vec{X}_{i_n} + \frac{\lambda}{M}\sum_{j=0}^{m}(\vec{X}_{j_n} - \vec{X}_{i_n}) \quad (5)$$

where n represents the number of iterations, i is the vertex index, $\vec{x}_{i_n}$ represents the deformation vector of the *i-th* vertex at the *n-th* iteration, $j = 0,1, \dots, m$ represents *m+1* indexes of neighbor vertices around vertex i, λ represents the smooth intensity coefficient.

In terms of the SAE neural network parameters set, we try to select the appropriate number of nodes in the hidden layer, and output layer as well as the number of levels in the SAE neural network by changing different parameter values in this study. After that, we determine to use a three-level SAE neural network in this study. The number of input nodes of the first level encoder is 752 × 3 = 2256 (i.e., the number of points of down-sampled mesh multiplied by the spatial dimension), and the corresponding number of output nodes is 1000. The number of input and output nodes of the second level encoder is 1000 and 200, respectively. The number of input and output nodes of the third level encoder is 200 and 30, respectively. Finally, the shape model including multiple organs is represented by the 30-dimensional feature vector $\vec{f}_3$, which is the output of the third level encoder.

3) Shape fitting and generation with SAE neural network

Based on the trained three-level SAE neural network, we can do multi-organ shape fitting and generation for new mouse micro-CT images. As shown in Fig. 3 (a), the shape vectors of multiple organs are input to a three-level SAE neural network. Through the process of three compressions and three decompressions, a new shape is reconstructed as the fitting result of the input shape. According to the principle of the SAE neural network, the output shape should be as similar as possible to the input shape, and the error between them indicates the fitting error. The smaller the fitting error is, the better the training performance of the SAE neural network is. In this paper, we calculate the average surface distance [20]

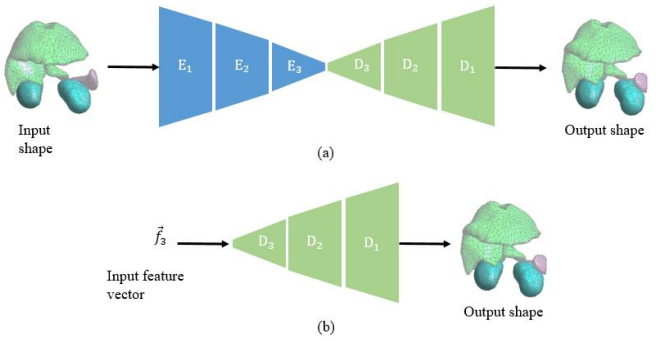

Fig. 3. Shape fitting and shape generation using the SAE neural network. (a) Shape fitting; (b) Shape generation.

between the constructed model and a training mesh to evaluate the modeling performance [21] of our method. If an input shape is one of the training samples, the fitting error reflects the specificity of the SAE neural network; if an input shape is not included in the training sample set, the fitting error reflects the generalization of the SAE neural network.

Fig. 3 (b) shows the shape generation process based on the three-level decoder. If a randomly selected feature vector $\vec{f}_3$ is input to the network, the corresponding output shape can be obtained; if the training performance of the network is effective, the output shape should correspond to the true anatomical shape. Further, if the mean shape of the training meshes is input to the three-level SAE neural network, the feature vector $\overline{\vec{f}_3}$ corresponding to the mean shape can be obtained. And if any value in any dimension of $\overline{\vec{f}_3}$ is modified, the deformation effect of the mean shape can be constructed, thus realizing the construction of the multi-resolution shape model for multiple organs.

III. Results

A. *Multi-resolution Shape Modeling Results Based on SAE Neural Network*

According to the shape generation method in the previous section, we take the 30-dimensional feature vector $\overline{\vec{f}_3}$ of the mean shape as the basis and make the unit size, which deviates from the average value, to be 0.1 by changing the value of each dimension in the study. At last, deformation results of the mean shape corresponding to 30 dimensions are obtained. we have found that in the process of controlling the deformation of the 30-dimensional components, some of the deformation laws of the model show a certain similarity. Therefore, in Fig. 4, 17 representative deformation components are selected, where $F_i(i = 1,2,3,4,6,7,8,11,13,14,15,16,18,19,25,26,30)$ represents the *i-th* deformation component of the third level in the SAE neural network. The first, third, and fifth columns are the mean models, and the second, fourth, and sixth columns are the deformation models reconstructed by adjusting these deformation components in this network. Some areas of the reconstruction model that have obvious deformation are circled. As Fig. 4 shows, in the reconstruction of the multi-organ shape model, different deformation components in the third level of the SAE neural network control the deformation of different areas of the model. Such as

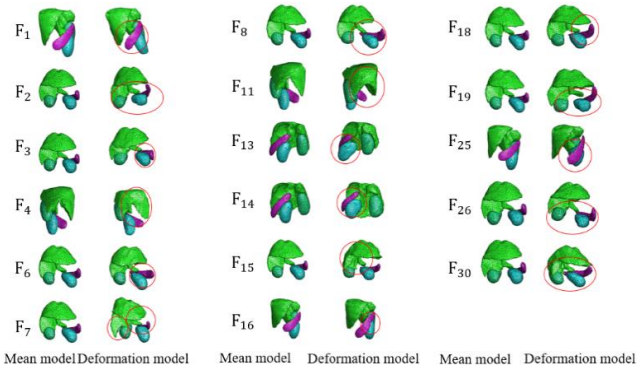

Fig. 4. Multi-organ model reconstructed by SAE neural network.

components F_4, F_7, F_{11}, F_{15} can control liver deformation, components F_2, F_3, F_6, F_{13} can control spleen deformation, components F_7, F_{15}, F_{30} can control right kidney deformation, components F_{13}, F_{14}, F_{25}, F_{30} can control left kidney deformation, and the components F_8, F_{26}, etc. can control the relative position changes of certain organs. Each deformation component of constructing a multi-organ model not only controls the deformation of a certain organ, but also affects the changes in the shape and position of multiple organs. In the constructed model, there is not only a single organ with obvious local deformations, but also multiple organs with obvious global deformations, which shows a multi-resolution deformation effect.

B. *Comparison of SAE Neural Network and SSM Modeling*

In the previous study, we used a modified traditional PCA method to construct traditional global SSM and multi-resolution SSM for the mouse micro-CT images and obtained global and local deformation modes [16]. On this basis, we further compare the model constructed by the SAE network in this study with the multi-resolution model constructed in the previous study. We quantitatively analyze the generalization and specificity of the models constructed by these two different approaches, based on the average surface distance errors of the constructed model and each training data.

For comparison, Fig. 5 shows the traditional PCA method to construct a global SSM for realizing the global deformation effect. The first and fourth columns are the mean models, which are represented by $\boldsymbol{\mu}$. The deformation components are represented by PC_1, PC_2 and PC_3. $\lambda_i(i = 1,2,\dots,10)$ is the eigenvalue obtained by implementing eigenvalue decomposition on the covariance matrix of training sample points, which corresponds to the deformation component. $\alpha_i(i = 1,2,\dots,10)$ is the model coefficient, which is set as the weight of the deformation model. Comparing Fig. 4 and Fig. 5, we can find that almost all the deformation models constructed by PCA methods can also be constructed by using SAE neural network. However, the multi-organ deformation model reconstructed by SAE neural network has more deformation modes than that constructed by PCA methods.

In this study, the SAE neural network is mainly used to reconstruct the mouse abdomen multi-organ model. As a comparison, the traditional global SSM and multi-resolution SSM are introduced to further analyze the accuracy of these

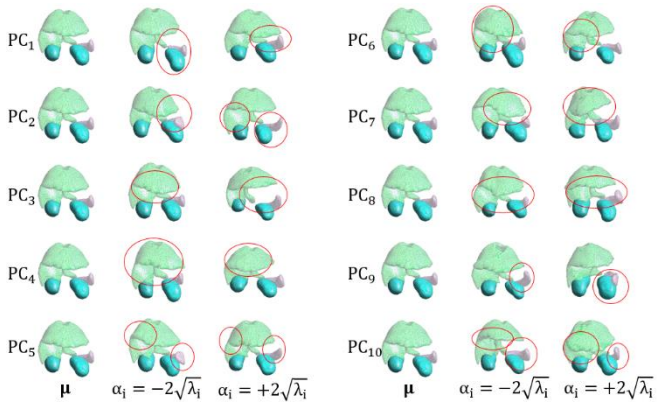

Fig. 5. Multi-organ model constructed by PCA method.

three modeling methods. It can be seen from Fig. 6 (a) that in terms of generalization, the modeling error of the traditional global multi-organ SSM and multi-resolution model are smaller than that of the SAE neural network. Especially for the multi-resolution model, the error is around 0.3 mm, while the modeling error of the SAE neural network is about 1.5 mm. In Fig. 6 (b), in terms of specificity, the modeling errors of the traditional global multi-organ SSM and the multi-resolution model have reached the level of 0.3 mm and 0.2 mm, respectively, but the modeling error of the SAE neural network is still around 1.5 mm. According to the analysis, although the principle of the SAE neural network is simple and easy to implement, it is still inferior to the linear SSM construction method based on the PCA method. We also find that the number of hidden layers, the number of nodes, and the sparseness of the SAE neural network may cause the training error to become larger, which needs to be verified in future work.

IV. Conclusion

The method proposed in this paper is more concise and feasible, and the model constructed by it can express much richer deformation modes. The SAE neural network is simpler than PCA-based methods in terms of algorithm complexity. But the deformation successfully reflects multi-resolution changes in shapes. Moreover, SAE neural network is a new type of multi-resolution modeling method based on shape prior knowledge and is easy to implement. SAE neural network has potential application for multi-resolution organ segmentation as the multi-resolution features are important for the shape modeling of multiple organs. The advantages of the proposed method are that the constructed diverse deformation components consist of nonlinear and linear modeling characteristics. For example, the deformation components obtained by SAE neural network are much more than that of PCA-based methods. Likely, the model built by SAE neural network consists of much more local deformation components which can change the local shape of an organ in detail. Although the error rates of generalization and specificity of the SAE neural network are still a little bit higher, the excellent nonlinear multi-resolution modeling characteristics deserve further improvement and optimization. On the other hand, SAE neural network is a nonlinear modeling method, which is more suitable for simulating nonlinear shape space with a manifold distribution. In theory, it can obtain a more accurate modeling effect than traditional linear modeling methods. Furthermore, SAE neural network modeling method has multi-resolution characteristics not shared by traditional methods, which means SAE neural network modeling method performs better theoretically.

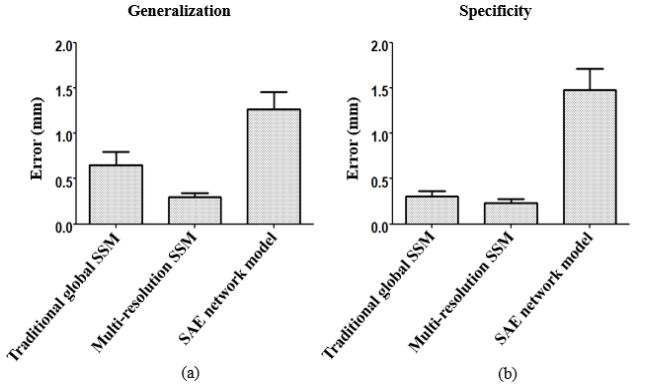

Fig. 6. Shape reconstruction accuracy comparison among the SAE neural network with the traditional and multi-resolution models. (a) Generalization errors; (b) Specificity errors.

Acknowledgment

We thank the Molecular Imaging Center of the University of California, Los Angeles for providing 98 mouse micro-CT images to support our work.

References

[1] S. J. Lim, Y. Y. Jeong, and Y. S. Ho, "Automatic liver segmentation for volume measurement in CT Images," *Journal of Visual Communication and Image Representation*. vol. 17, pp. 860–875, 2006.

[2] N. Karunanayake, P. Aimmanee, W. Lohitvisate, and S.S. Makhanov, "Particle method for segmentation of breast tumors in ultrasound images," *Mathematics and Computers in Simulation*. vol. 170, pp. 257–284, 2020.

[3] A. Klein, J. Warszawski, J. Hillengaß, and K. H. Maier-Hein, "Automatic bone segmentation in whole-body CT images," *International Journal of Computer Assisted Radiology and Surgery*. vol. 14, pp. 21–29, 2019.

[4] X. Zhuang, K. S. Rhode, R. S. Razavi, D. J. Hawkes, and S. Ourselin, "A registration-based propagation framework for automatic whole heart segmentation of cardiac MRI," *IEEE Transactions on Medical Imaging*. vol. 29, pp. 1612–1625, September 2010.

[5] M. B. Stegmann and D. D. Gomez, "A brief introduction to statistical shape analysis," *Informatics and Mathematical Modelling, Technical University of Denmark*, Denmark, pp. 1–15, March, 2002.

[6] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models – their training and application," *Computer Vision and Image Understanding*. vol. 61, pp. 38–59, 1995.

[7] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*. vol. 23, pp. 681–685, June 2001.

[8] P. Tresadern, H. Bhaskar, S. A. Adeshina, and C. J. Taylor, "Combining local and global shape models for deformable object matching," *British Machine Vision Conference*. London: BMVA Press, 2009, pp. 1–12.

[9] M. Wilms, H. handels, and J. Ehrhardt, "Multi-resolution multi-object statistical shape models based on the locality assumption," *Medical Image Analysis*. vol. 38, pp. 17–29, 2017.

[10] P. D. Sozou, T. F. Cootes, C. J. Taylor, and E. C. Di-Mauro, "A non-linear generalisation of PDMs using polynomial regression," *British Machine Vision Conference*. York : BMVA Press, 1994, pp. 397–406.

[11] P. D. Sozou, T. F. Cootes, C. J. Taylor, E. C. Di Mauro, and A. Lanitis, “Non-linear point distribution modelling using a multi-layer perceptron,” *Image and Vision Computing*. vol. 15, pp. 457–463, 1997.

[12] C. J. Twining and C. J. Taylor, “Kernel principal component analysis and the construction of non-linear active shape models,” *British Machine Vision Conference*. Manchester: BMVA Press, 2001, pp. 1–10.

[13] H. Huang, F. Makedon, and R. McColl, “High dimensional statistical shape model for medical image analysis,” *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. pp. 1541–1544, 2008.

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

[15] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black, “Generating 3D faces using convolutional mesh autoencoders,” *European Conference on Computer Vision*. Cham: Springer, 2018, pp. 725–741.

[16] Z. Chen, T. Ristaniemi, F. Cong, and H. Wang, “Multi-resolution statistical shape models for multi-organ shape modelling,” *International Symposium on Neural Networks*. Cham: Springer, 2020, pp. 74–84.

[17] W. E. Lorensen and H. E. Cline, “Marching Cubes: A high resolution 3D surface construction algorithm,” *Computer Graphics*. vol. 21, pp. 163–169, July 1987.

[18] R. Marani, V. Renò, M. Nitti, T. D'Orazio, and E. Stella, “A modified iterative closest point algorithm for 3D point cloud registration,” *Computer-Aided Civil and Infrastructure Engineering*. vol. 31, pp. 515–534, July 2016.

[19] S. I. Park and S.-J. Lim, “Template-based reconstruction of surface mesh animation from point cloud animation,” *ETRI Journal*. vol. 36, pp. 1008–1015, December 2014.

[20] H. Wang, D. B. Stout, R. Taschereau, Z. Gu, N. T. Vu, and D. L. Prout et al, “MARS: a mouse atlas registration system based on a planar x-ray projector and an optical camera,” *Physics in Medicine and Biology*. vol. 57, pp. 6063–6077, October 2012.

[21] T. Heimann and H. P. Meinzer, “Statistical shape models for 3D medical image segmentation: a review,” *Medical Image Analysis*. vol. 13, pp. 543–563, August 2009.