Analysis of Generative Shape Modeling Approaches: Latent Space Properties and Interpretability

Hristina Uzunova[1], Jesse Kruse[1], Paul Kaftan[1], Matthias Wilms[2],
Nils D. Forkert[2], Heinz Handels[1], Jan Ehrhardt[1]

[1]Institute of Medical Informatics, University of Lübeck, Germany
[2]Department of Radiology, University of Calgary, Canada
uzunova@imi.uni-luebeck.de

Abstract. Generative shape models are crucial for many medical image analysis tasks. In previous studies, it has been shown that conventional methods like PCA-based statistical shape models (SSMs) and their extensions are thought to be robust in terms of generalization ability but have rather poor specificity. On the contrary, deep learning approaches like autoencoders, require large training set sizes, but are comparably specific. In this work, we comprehensively compare different classical and deep learning-based generative shape modeling approaches and demonstrate their limitations and advantages. Experiments on a publicly available 2D chest X-ray data set show that the deep learning methods achieve better specificity and similar generalization abilities for large training set sizes. Furthermore, an extensive analysis of the different methods, gives an insight on their latent space representations.

1 Introduction

The study of anatomical shapes is a fundamental process in medical image analysis. Generative shape modeling methods seek to capture as much information as possible to estimate the shape distribution of a population. Typically, a training set of shapes is used to train a model that is able to reproduce the training shapes and to generate new but similar shape instances. Applications of these models range from segmentation and registration, over data augmentation, to the detection and classification of diseases in medical images. A classical example of such generative models are PCA-based statistical shape models (SSMs) introduced by Cootes et al. in the early 1990s [1]. Since then, numerous extensions and modifications of these models have been proposed to alleviate problems and limitations regarding the linear nature of this approach [2], the need for 1-to-1 correspondences [3], or reduce the amount of training data required [4,5]. More recently, the research focus has shifted towards deep learning-based generative modeling using approaches like autoencoders (AEs), variational autoencoders (VAEs), or generative adversarial networks (GANs). Although these approaches overcome major limitations of SSMs, they require even larger training data sets and their black-box nature makes them harder to interpret. Nevertheless, deep learning methods are successful, especially for the generation of synthetic images [6]. However, in the medical field training data is limited and only a few studies have examined the properties of generative models for this specific situation [7].

In a previous study [8], two conventional and two deep learning-based generative approaches for multi-organ shape modeling were compared. That study showed that approaches based on deep learning consistently show better results in terms of generalizability and specificity than classical PCA-based SSMs. Yet, it has been shown that extensions of classical SSMs [5] perform on par with and even outperform deep learning- based approaches in terms of generalizability, especially for smaller training sets. Moreover, deep learning approaches tend to model small or rare structures incorrectly when only few samples are available.

In this work, we extend [8] by (1) including other popular generative models, (2) investigating solutions for the incorrect generation of small structures in (V)AEs, and (3) examining the properties of the latent spaces of the models.

2 Materials and Methods

2.1 Material

The generative shape models discussed in this paper use a publicly available 2D chest radiograph database [9]. We refrain from using 3D data here to avoid typical computational problems that would require special solutions for the deep learning approaches. The dataset contains 247 images with segmentations of five structures (left and right lung, left and right clavicle, and the heart). Segmentations are given as a set of 166 corresponding landmarks (input for the SSMs) and as binarized label images (input for the deep learning approaches; see Fig. 1). We use the same fixed 123/124 images test/training split as in [9] to allow for a direct comparison of the results.

2.2 Statistical Shape Models and Locality-based Multi-resolution SSMs

Statistical Shape Models (SSMs) use vectorized representations of landmark points that represent the shape of the object, typically the object contours. Principal Component Analysis (PCA) of a training set is used to create an

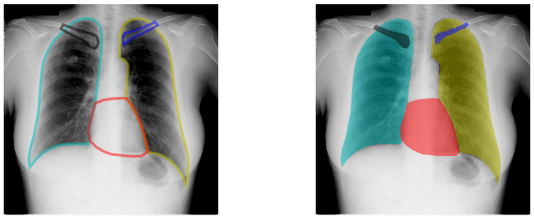


Fig. 1. Example shapes as contours (left) for the SSMs; and labels (right) for the CNNs.

orthonormal basis for projecting shape representations into a low-dimensional latent space or to reconstruct new shapes from latent representations [1]. In classical SSMs, the number of training samples influences the flexibility of the model, since the size of the latent space is limited by the size of the training set. *Locality-based multi-resolution SSMs (LSSMs)* [5] introduce additional flexibility by breaking global relationships and assuming that local shape variations have limited effects in distant areas. This idea can be integrated into the traditional SSM framework by manipulating covariances based on the distance between landmarks in a multi-resolution manner. LSSMs have been shown to perform on par or outperform other approaches like wavelet-based SSMs or Gaussian process models in terms of generalization and specificity [5,10].

2.3 Autoencoders and Variational Autoencoders

Autoencoders (AEs) are neural networks consisting of an encoder $Q(X)$ that maps input data X to a low-dimensional latent vector $\mathbf{z}$, and a decoder $P(\mathbf{z})$ that attempts to reconstruct the input data X given $\mathbf{z}$. This is typically achieved by optimizing a reconstruction objective, s.t. $X \approx P(Q(X))$. Thus, unseen shapes can be reconstructed by forwarding them through a trained encoder and decoder. To generate new shapes, a random $\mathbf{z}$ can be sampled and propagated through a trained decoder [8].

However, AEs may simply learn an identity function and because of the unknown latent distribution, sampling a random $\mathbf{z}$ can cause the generation of implausible shapes. Those problems are addressed by *variational autoencoders (VAEs)*, by restricting the latent space to a normal distribution. In practice, this is achieved by an additional Kullback-Leibler loss of the latent space [11].

2.4 Generative Adversarial Networks

Generative adversarial networks (GANs) can analogously be used as generative models. However, due to their adversarial training scheme, they are known to enable the generation of exceptionally realistic images. GANs learn to map a random noise vector $\mathbf{z}$ to an output image X_{fake} using a generator function $G : \mathbf{z} \to X_{fake}$ [6]. To ensure that the generator produces realistic images, an adversarial discriminator D is used during training, aiming at perfectly distinguishing real images and generated fake data.

2.5 Deformable Autoencoders

AEs or VAEs often fail to reconstruct or generate small-sized structures, especially when trained on small datasets [8]. *Deformable autoencoders (DAE)* are an extension of VAEs that tackle this problem by representing shapes as the deformed version of a learned template image [12]. Rather than directly reconstructing the input, the DAE decoder generates a displacement field φ and implicitly learns a template T, which is deformed to match the input X s.t. $T \circ P(z) \approx X$. To ensure smooth displacement fields, a diffusion regularisation term is used as an additional penalty during model training.

3 Experiments and Results

In this work, we focus on analyzing differences between the approaches presented in Sec. 2 with respect to specificity and generalization and also investigate the structure of each model's latent space. The experimental setup follows [8].

Specificity and Generalization: In this experiment, all methods are compared in terms of generalization and specificity for different training set sizes, where specificity describes the model's ability to generate new realistic samples and generalization denotes the ability to reconstruct unseen samples. As in [8], training set sizes ranging from 5 to 113 are used for a 5-fold-cross-validation. Average symmetric surface/contour distances (ASSD) are utilized to quantify the results.

The results for all methods are shown in Fig. 2. The deep learning methods show overall better results in terms of specificity. However, for small training set sizes (< 60), the LSSM model shows improved generalization abilities. For large training set sizes, the generalization abilities of the methods are roughly the same. While the deep learning methods seem to have similar generalization abilities, the DAE and GAN models achieve the best specificity.

Furthermore, for smaller training set sizes, VAEs and AEs are not able to reconstruct all labels, especially small labels like the clavicles. Thus, the percentage of images missing one or more labels is fairly high. This problem does not appear for the DAE model. However, the DAE seems to require larger training datasets to generate a proper template and consequently achieve a competitive generalization performance (see Fig. 3).

Latent Space: The latent space of the methods presented here is crucial to their representation ability. The latent space size of the traditional methods lies in the ranges $[3, 14]$ (SSM) and $[4, 55]$ (LSSM) depending on the training set size. Compared to the deep learning methods with a fixed latent space of size 512, the traditional methods have fairly compact latent spaces contributing to their interpretability. To further investigate the structure of the latent spaces, we linearly interpolate between the latent vectors of two randomly chosen images and project the interpolated vectors back to image space. In Fig. 4, examples for AE and DAE show that smooth interpolations are possible. However, AEs tend to generate artifacts and implausible shapes for smaller training sets. This is avoided in DAEs due to applying smooth deformations to a generated template.

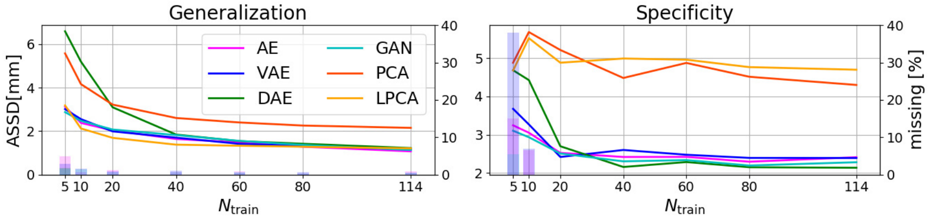


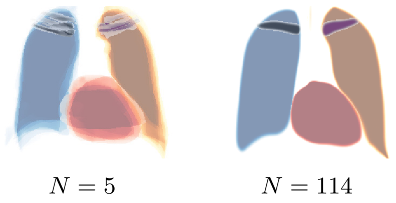
Fig. 2. Generalization and specificity of the models for varying training set sizes (lower values are better). The bars indicate the percentage of images with missing labels.

Typically, a normal distribution is assumed for the latent space of the models when sampling new shapes. To verify this assumption, we apply a component-wise Shapiro-Wilk normality test on the latent encodings of the training and test data and calculate the percentage of non-normally distributed components (Fig. 3). Due to the explicit normalization of the latent spaces of the VAE and DAE, they have a small percentage of non-normally distributed components ($\sim 10\% - 15\%$). This value increases up to 40% for PCA and LPCA due to their linear nature, accounting for their worse specificity.

A further important property is that, in contrast to SSMs, neural networks do not guarantee that a reconstructed image is mapped to the same latent vector as the original image. We, therefore, calculate the distance between the encoding of the input image and the encoding of the reconstructed image averaged over all training data. To establish a consequent scheme, a Mahalanobis distance was used for all methods. As baseline (BL), we approximate the mean pairwise distance of the latent encodings. The values obtained are: VAE 0.01 (BL 0.41); AE 11 (BL 26.7); GAN 11 (BL 21.8); DAE 0.02 (BL 0.6). Those distances indicate an ambiguity and poor interpretability of the latent space, that is not present for the classical shape models.

4 Discussion and Conclusion

In this work, we compared the shape modelling abilities of two classical and four deep learning methods. In terms of generalization, the classical LPCA method is more robust for small training set sizes, whereas for large training set sizes, it is on par with the deep learning methods. Most deep learning methods fail to generate small structures when trained on small datasets, while deformable autoencoders cope with this problem, yet they require a considerable amount of training data to reach the generalization ability of the other approaches. In terms of specificity, the deep learning methods show significantly better results. The latent spaces of SSMs are much more compact and intuitive, however, their linear nature yields non-normally distributed latent spaces explaining their poor specificity. An interpolation experiment visualizes the smoothness of the latent


$N = 5$ $N = 114$

method	train data	test data
VAE	10.1%	9.8%
DAE	15.1%	12.7%
GAN	19.7%	18.6%
AE	19.7%	18.0%
PCA	28.8%	32.7%
LPCA	27.9%	39.5%

Fig. 3. Left: The DAE learned templates with a different amount of training images. Right: Percent of non-normally distributed components of the latent vectors of all models determined by a Shapiro-Wilk test.

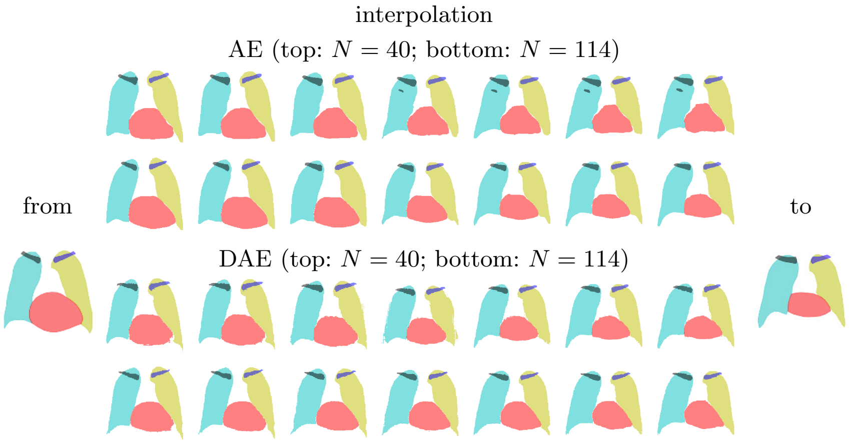

interpolation
AE (top: $N = 40$; bottom: $N = 114$)
from
to
DAE (top: $N = 40$; bottom: $N = 114$)

Fig. 4. Projected latent space interpolation between two shapes for different methods.

spaces and shows the strengths of DAEs compared to AEs. Still, a rather large drawback of the deep learning methods is the ambiguity of their latent space.

References

1. Cootes TF, Taylor CJ, Cooper DH, et al. Active Shape Models-Their Training and Application. Comput Vis Image Underst. 1995;61(1):38–59.
2. Kirschner M, Becker M, Wesarg S. 3D Active Shape Model Segmentation with Nonlinear Shape Priors. In: MICCAI; 2011. p. 492–499.
3. Krüger J, Ehrhardt J, Handels H. Statistical Appearance Models Based on Probabilistic Correspondences. Med Image Anal. 2017;37:146–159.
4. Davatzikos C, Tao X, Shen D. Hierarchical Active Shape Models, Using the Wavelet Transform. IEEE Trans Med Imaging; p. 2003.
5. Wilms M, Handels H, Ehrhardt J. Multi-Resolution Multi-Object Statistical Shape Models Based on the Locality Assumption. Med Image Anal. 2017;38:17–29.
6. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Nets. In: Advances in Neural Information Processing Systems; 2014. p. 2672–2680.
7. Ruan X, Murphy RF. Evaluation of methods for generative modeling of cell and nuclear shape. Bioinformatics. 2018 12;35(14):2475–2485.
8. Uzunova H, Kaftan P, Wilms M, et al. Quantitative Comparison of Generative Shape Models for Medical Images. In: BVM; 2020. p. 201–207.
9. van Ginneken B, Stegmann MB, Loog M. Segmentation of Anatomical Structures in Chest Radiographs Using Supervised Methods. Med Image Anal. 2006; p. 19–40.
10. Wilms M, Ehrhardt J, Forkert ND. A Kernelized Multi-Level Localization Method for Flexible Shape Modeling with Few Training Data. In: MICCAI; 2020. p. 765–775.
11. Kingma D, Welling M. Auto-Encoding Variational Bayes. In: International Conference on Learning Representations; 2014. .
12. Shu Z, Sahasrabudhe M, Alp Güler R, et al. Deforming Autoencoders: Unsupervised Disentangling of Shape and Appearance. In: ECCV; 2018. p. 664–680.