Uncertain-DeepSSM: From Images to Probabilistic Shape Models

Jadie Adams[1,2], **Riddhish Bhalodia**[1,2], **Shireen Elhabian**[1,2]

[1]Scientific Computing and Imaging Institute, University of Utah, UT, USA

[2]School of Computing, University of Utah, UT, USA

Abstract

Statistical shape modeling (SSM) has recently taken advantage of advances in deep learning to alleviate the need for a time-consuming and expert-driven workflow of anatomy segmentation, shape registration, and the optimization of population-level shape representations. DeepSSM is an end-to-end deep learning approach that extracts statistical shape representation directly from unsegmented images with little manual overhead. It performs comparably with state-of-the-art shape modeling methods for estimating morphologies that are viable for subsequent downstream tasks. Nonetheless, DeepSSM produces an overconfident estimate of shape that cannot be blindly assumed to be accurate. Hence, conveying what DeepSSM does not know, via quantifying granular estimates of uncertainty, is critical for its direct clinical application as an on-demand diagnostic tool to determine how trustworthy the model output is. Here, we propose Uncertain-DeepSSM as a unified model that quantifies both, data-dependent aleatoric uncertainty by adapting the network to predict intrinsic input variance, and model-dependent epistemic uncertainty via a Monte Carlo dropout sampling to approximate a variational distribution over the network parameters. Experiments show an accuracy improvement over DeepSSM while maintaining the same benefits of being end-to-end with little pre-processing.

Keywords

Uncertainty Quantification; Statistical Shape Modeling; Bayesian Deep Learning

1 Introduction

Morphometrics and its new generation, statistical shape modeling (SSM), have evolved into an indispensable tool in medical and biological sciences to study anatomical forms. SSM has enabled a wide range of biomedical and clinical applications (e.g., [2, 4, 6, 7, 17–19, 22, 31, 47, 49]). Morphological analysis requires parsing the anatomy into a quantitative representation consistent across the population at hand to facilitate the testing of biologically relevant hypotheses. A popular choice for such a representation is using *landmarks* that are defined consistently using invariant points, i.e., *correspondences*, across populations [43]. Coordinate transformations (e.g., [26, 27]) hold promise as an alternative representation, but the challenge is finding the anatomically-relevant transformation that quantifies differences among shapes. Ideally, landmarking is performed by anatomy experts to mark distinct, and typically few anatomical features [1, 33], but it is time-intensive and cost-prohibitive,

especially for 3D images and large cohorts. More recently, dense sets of correspondence points that capture population statistics are used, thanks to advances in computationally driven approaches for shape modeling (e.g., [8, 9, 11, 13, 45]).

Traditional computational approaches to automatically generate dense correspondence models, a.k.a. point distribution models (PDMs), still entail a time-consuming, expert-driven, and error-prone workflow of segmenting anatomies from volumetric images, followed by a processing pipeline of shape registration, correspondence optimization, and projecting points onto some low-dimensional shape space for subsequent statistical analysis. Many of these steps require significant parameter tuning and/or quality control by the users. The excessive time and effort to construct population-specific shape models have motivated the use of deep networks and their inherent ability to learn complex functional mappings to regress shape information directly from images and incorporate prior knowledge of shapes in image segmentation tasks (e.g., [3, 25, 38, 48, 50]). However, deep learning in this context has drawbacks. Training deep networks on volumetric images is often confounded by the combination of high-dimensional image spaces and limited availability of training images labeled with shape information. Additionally, deep networks can make poor predictions with no indication of uncertainty when the training data weakly represents the input. Computationally efficient automated morphology assessment when integrated with new clinical tools as well as surgical procedure, has potential to improve medical care standards and clinical decision making. However, uncertainty quantification is a must in such scenarios, as it will allow professionals to determine the trustworthiness of such a tool and prevent unsafe predictions.

Here, we focus on a particular instance of a deep learning-based framework, namely DeepSSM [3], that maps unsegmented 3D images to a low-dimensional shape descriptor. Mapping to a low-dimensional manifold, compared with regressing correspondence points, has a regularization effect that compensates for misleading image information and provides a topology-preserving prior to shape estimation. DeepSSM also entails a population-driven data augmentation approach that addresses limited training data, which is typical with small and large-scale shape variability. DeepSSM has been proven effective in characterizing pathology [3] and performs statistically similar to traditional SSM methods in downstream tasks such as disease recurrence predictions [4]. Nonetheless, DeepSSM, like other deep learning-based frameworks predicting shape, produces an overconfident estimate of shape that can not be blindly assumed to be accurate. Furthermore, the statistic-preserving data augmentation is bounded by what the finite set of training samples captures about the underlying data distribution. In this paper, we formalize Uncertain-DeepSSM, a unified solution to limited training images with dense correspondences and model prediction over-confidence. Uncertain-DeepSSM quantifies granular estimates of uncertainty with respect to a low-dimensional shape descriptor to provide spatiallycoherent, localized uncertainty measures (see Fig. 1) that are robust to misconstruing factors that would typically affect point-wise regression, such as heterogeneous image intensities and noisy diffuse shape boundaries. Uncertain-DeepSSM produces probabilistic shape models directly from 3D images, conveying what DeepSSM does not know about the input and providing an accuracy improvement over DeepSSM while being end-to-end with little required pre-processing.

2 Related Work

DeepSSM [3] is based on works that show the efficacy of convolutional neural networks (CNNs) to extract shape information from images. Huang *et al.* [25] regress shape orientation and position conditioned on 2D ultrasound images. Milletari *et al.* [38] segment ultrasound images using low-dimensional shape representation in the form of principal component analysis (PCA) to regress landmark positions. Oktay *et al.* [41] incorporate prior knowledge about organ shape and location into a deep network to anatomically constrain resulting segmentations. However, these works provide a point-estimate solution for the task at hand.

In Bayesian modeling, there are two types of uncertainties [28, 30]. *Aleatoric* (or data) uncertainty captures the uncertainty inherent in the input data, such as over-exposure, noise, and the lack of the image-based features indicative of shapes. *Epistemic* (or model) uncertainty accounts for uncertainty in the model parameters and can be explained away, given enough training data [28].

Aleatoric uncertainty can be captured by placing a distribution over the model output. In image segmentation tasks, this has been achieved by sampling segmentations from an estimated posterior distribution [10, 34] and using conditional normalizing flows [44] to infer a distribution of plausible segmentations conditioned on the input image. These efforts succeed in providing shape segmentation with aleatoric uncertainty measures, but do not provide a shape representation that can be readily used for population-level statistical analyses. Tóthová *et al.* [46] incorporate prior shape information into a deep network in the form of a PCA model to reconstruct surfaces from 2D images with an aleatoric uncertainty measure that is quantified via conditional probability estimation. Besides being limited to 2D images, quantifying point-wise aleatoric uncertainty makes this measure prone to inherent noise in images.

Epistemic uncertainty is more difficult to model as it requires placing distributions over models rather than their output. Bayesian neural networks [12, 14, 37] achieve this by placing a prior over the model parameters, then quantifying their variability. Monte Carlo dropout sampling, which places a Bernoulli distribution over model parameters [15], has effectively been formalized as a Bayesian approximation for capturing epistemic uncertainty [16]. Aleatoric and epistemic uncertainty measures have been combined in one model for tasks such as semantic segmentation, depth regression, classification, and image translation [28, 32, 42], but never for SSM.

Uncertain-DeepSSM produces probabilistic shape models directly from images that quantifies both the data-dependent aleatoric uncertainty and the model-dependent epistemic uncertainty. We quantify aleatoric uncertainty by adapting the network to predict intrinsic input variance in the form of mean and variance for the PCA scores and updating the loss function accordingly [35, 40]. This enables explicit modeling of the heteroscedastic-type of aleatoric uncertainty, which is dependent on the input data. We model epistemic uncertainty via a Monte Carlo dropout sampling to approximate a variational distribution over the network parameters by sampling PCA score predictions with various dropout masks. This

approach provides both uncertainty measures for each PCA score that are then mapped back to the shape space for interpretable visualization. Uncertainty fields on estimated 3D shapes convey insights for how the given input relate to what the model knows. For instance, such uncertainties could help pre-screen for pathology if Uncertain-DeepSSM is trained on controls. Furthermore, explicit modeling of uncertainties in Uncertain-DeepSSM provides more accurate predictions, compared with DeepSSM, with no additional training steps. This indicates the ability of Uncertain-DeepSSM to better generalize in limited training data setting.

3 Methods

A trained Uncertain-DeepSSM model provides shape descriptors, specifically PCA scores, with uncertainty measures directly from 3D images (e.g., CT, MRI) of anatomies. In this section, we describe the data augmentation method, the network architecture, training strategy, and uncertainty quantification.

3.1 Notations

Consider a paired dataset $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$ of N 3D images $\mathbf{y}_n \in \mathbb{R}^{H \times W \times D}$ and their corresponding shapes $\mathbf{x}_n \in \mathbb{R}^{3M}$, where each shape is represented by M 3D correspondence points. We generate a PDM from segmentations. This entails the typical SSM pipeline that includes pre-processing steps (registration, resampling, smoothing, …), and correspondence (i.e., PDM) optimization. In practice, any PDM generation algorithm can be employed. Here, we use the open-source *ShapeWorks* software [8] to optimize surface correspondences using anatomies segmented from the *training* images. Next, high-dimensional shapes (i.e., PDM) in the shape space (of dimension $\mathbb{R}^{3M}$) are mapped to low-dimensional PCA scores $\mathbf{z} \in \mathbb{R}^L$ in the PCA subspace that is parameterized by a mean vector $\boldsymbol{\mu} \in \mathbb{R}^{3M}$, a diagonal matrix of eigen values $\boldsymbol{\Delta} \in \mathbb{R}^{L \times L}$, and a matrix of eigen vectors $\mathbf{U} \in \mathbb{R}^{3M \times L}$, where $\mathbf{z} = \mathbf{U}^T(\mathbf{x} - \boldsymbol{\mu})$ and $L \ll 3M$ is chosen such that at least 95% of the population variation is explained. The PCA scores $\mathbf{z}_n$ associated with the training image $\mathbf{y}_n$ serve as a supervised target to be inferred by the Uncertain-DeepSSM network and mapped deterministically to correspondence points, where $\mathbf{x}_n = \mathbf{U}\mathbf{z}_n + \boldsymbol{\mu}$. The network thus defines a functional map $f_{\boldsymbol{\Theta}}: \mathbb{R}^{H \times W \times D} \to \mathbb{R}^L$ that is parameterized by the network parameters $\boldsymbol{\Theta}$, where $\mathbf{z} = f_{\boldsymbol{\Theta}}(\mathbf{y})$. Uncertainties are quantified in the PCA subspace, such that the PCA scores of the n–th training shape $\mathbf{z}_n$ is associated with vectors of aleatoric variances $\mathbf{a}_n \in \mathbb{R}_+^L$ and epistemic variances $\mathbf{e}_n \in \mathbb{R}_+^L$.

3.2 Data Augmentation

DeepSSM augments training data with shape samples generated from a single multivariate Gaussian distribution in the PCA subspace, where an add-reject strategy is employed to prevent outliers from being sampled. Instead of assuming a Gaussian distribution, we use a kernel density estimate (KDE) to better capture the underlying shape distribution by not over-estimating the variance and avoiding sampling implausible shapes from high

probability regions in case of multi-modal distributions. Using KDE, augmented samples $\mathbf{z}_s$ are drawn from:

$$p_\sigma(\mathbf{z}) = \frac{1}{N} \sum_{n=1}^{N} K_n^\sigma(\mathbf{z}), \quad \text{s.t.} \quad K_n^\sigma(\mathbf{z}) = \frac{1}{\left(2\pi\sigma^2\right)^{L/2}} \exp\left(-\frac{\|\mathbf{z} - \mathbf{z}_n\|^2}{2\sigma^2}\right), \tag{1}$$

where $\sigma \in \mathbb{R}_+$ denotes the kernel bandwidth and is computed as the average nearest neighbor distance in the PCA subspace, i.e., $\sigma^2 = \frac{1}{N} \sum_n \min_{k \neq n}(\mathbf{z}_n - \mathbf{z}_k)^T \boldsymbol{\Delta}^{-1}(\mathbf{z}_n - \mathbf{z}_k)$. As illustrated in Fig. 2, a sampled vector of PCA scores $\mathbf{z}_s \in \mathbb{R}^L$ from the kernel of the n–th training sample $K_n^\sigma(\mathbf{z})$ is mapped to correspondence points $\mathbf{x}_s \in \mathbb{R}^{3M}$, where $\mathbf{x}_s = \mathbf{U}\mathbf{z}_s + \boldsymbol{\mu}$. Using the $\mathbf{x}_n \leftrightarrow \mathbf{x}_s$ correspondences, we compute thin-plate spline (TPS) warp [5] to obtain a deformation field which is then applied to image $\mathbf{y}_n$ to construct the augmented image $\mathbf{y}_s$. With this augmentation method, we can construct an augmented training set $\{(\mathbf{x}_s, \mathbf{y}_s)\}_{s=1}^{S}$ of S 3D images, their corresponding shapes, and the supervised targets $\{\mathbf{z}_s\}_{s=1}^{S}$, which respects the population-level shape statistics and the intensity profiles of the original dataset.

3.3 Adaptations for Uncertainty Quantification

We extend the network architecture and loss function of DeepSSM to estimate both types of uncertainties and the shape descriptor in the form of PCA scores.

Heteroscedastic aleatoric uncertainty is a measure of data uncertainty, and hence can be learned as a function of the input. Given a training set $\mathscr{D} = \{(\mathbf{y}_i, \mathbf{z}_i)\}_{i=1}^{I}$ that includes both real and augmented samples, where $I = N + S$, DeepSSM is trained to minimize the L2 loss between groundtruth $\mathbf{z}_i$ and predicted $f_{\boldsymbol{\Theta}}(\mathbf{y}_i)$. In Uncertain-DeepSSM, the network architecture is modified to estimate both the mean $\bar{\mathbf{z}}_i$ and variance $\mathbf{a}_i$ of the PCA scores, where $[\bar{\mathbf{z}}_i, \mathbf{a}_i] = f_{\boldsymbol{\Theta}}(\mathbf{y}_i)$. The variance acts as an uncertainty regularization term that does not require a supervised target since it is learned implicitly through supervising the regression task. For training purposes, we let the network predict the log of the variance, $\tilde{a}_{il} = \log a_{il}^2$, where a_{il} captures the aleatoric uncertainty along the l–th PCA mode of variation. This forces the variance to be positive and removes the potential for division by zero. Uncertain-DeepSSM is thus trained to minimize the Bayesian loss in (2), where $\bar{\mathbf{z}}_i = f_{\boldsymbol{\Theta}}^{\mathbf{z}}(\mathbf{y}_i)$ and $\tilde{\mathbf{a}}_i = f_{\boldsymbol{\Theta}}^{\mathbf{a}}(\mathbf{y}_i)$ are the $\mathbf{z}$– and $\mathbf{a}$– outputs of the network, respectively (see Fig. 3):

$$\mathscr{L}(\boldsymbol{\Theta}) = \frac{1}{2LI} \sum_{i=1}^{I} \left\{ [\mathbf{z}_i - \bar{\mathbf{z}}_i]^T \operatorname{diag}\left(\exp(\tilde{\mathbf{a}}_i)\right)^{-1} [\mathbf{z}_i - \bar{\mathbf{z}}_i] + \sum_{l=1}^{L} \tilde{a}_{il} \right\}. \tag{2}$$

The second term in (2) learns a loss attenuation, preventing the network from predicting infinite variance for all scores.

Epistemic uncertainty is a measure of the model's ignorance that can be quantified by modeling distributions over the model parameters $\boldsymbol{\Theta}$. We place a Bernoulli distribution over network weights by making use of the Monte Carlo dropout sampling technique [15, 16]. In

particular, a dropout layer with a probability κ is added to every layer (convolutional and fully connected) in the Uncertain-DeepSSM network (Fig. 3). Dropout is used in both, training and testing, where in testing, it is used to sample from the approximate posterior. The various dropout masks provide an ensemble of networks to sample predictions from, the distribution of which reflects the model's epistemic uncertainty. Consider V dropout samples, the epistemic uncertainty of the l–th PCA mode of variation is computed as,

$$e_{il} = \frac{1}{V}\sum_{v=1}^{V}\left(\bar{z}_{il}^{(v)}\right)^2 - \left(\frac{1}{V}\sum_{v=1}^{V}\bar{z}_{il}^{(v)}\right)^2 \tag{3}$$

where $\bar{\mathbf{z}}_{i}^{(v)} = f_{\boldsymbol{\Theta}_v \sim p(\boldsymbol{\Theta})}^{\mathbf{z}}(\mathbf{y}_i)$ is the $\mathbf{z}$–output of the network for the randomly masked network parameters $\boldsymbol{\Theta}_v$.

3.4 Architecture and Training

The network architecture of Uncertain-DeepSSM (Fig. 3) is similar to DeepSSM with five convolution layers followed by two fully connected layers. However in Uncertain-DeepSSM, dropout is added and batch normalization is removed. Combining batch normalization and dropout leads to a variance shift that causes training instability [36]. Hence, we normalize the input images to compensate for not using batch normalization. The PCA scores are also whitened to prevent the model from favoring the dominant PCA modes. A dropout layer with a probability of $\kappa = 0.2$ is added after every convolutional and fully connected layer. Data augmentation is used to create a set of $I = 4000$ training and 1000 validation images for training the network. PyTorch is used in constructing and training DeepSSM with Adam optimization [29] and a learning rate of 0.0001. Parametric ReLU [23] activation and Xavier weight initialization [21] are used. To train Uncertain-DeepSSM, the L2 loss function is used for the first epoch and the Bayesian loss function (2) is used for all following epochs. This allows the network to learn based on the task alone before learning to quantify uncertainty, resulting in better predictions and more stable training.

3.5 Testing and Uncertainty Analysis

When testing, dropout remains on and predictions are sampled multiple times. The predicted PCA scores and aleatoric uncertainty measure are first un-whitened. Using the dropout samples, the epistemic uncertainty measure is computed using (3) based on the un-whitened predicted scores. To compute the accuracy of the predictions, we first map PCA scores to correspondence points and compare the surface reconstructed from these points to the surface constructed from the ground truth segmentation. For surface reconstruction, we use the point correspondences between the population mean and the correspondence points of the predicted PCA scores to define a TPS warp that deform the mean mesh (from ShapeWorks [8]) to obtain the surface mesh for the predicted scores. The error is then calculated as the average of the surface-to-surface distance from the predicted to ground truth mesh and that of ground truth to predicted mesh.

To visualize uncertainty measures on the predicted mesh, the location of each correspondence point is modeled as a Gaussian distribution. To fit these distributions, we

sample PCA scores from a Gaussian with the predicted mean and desired variance (aleatoric or epistemic), then map them to the PDM space. This provides us with a distribution over each correspondence point, with mean and entropy indicating the coordinates of the point and the associated uncertainty scalar, respectively. Interpolation is then used to interpolate uncertainty scalars defined on correspondence points to the full reconstructed mesh.

4 Results

We compare the shape predictions from DeepSSM and Uncertain-DeepSSM on two 3D datasets; a toy dataset of parametric shapes (supershapes) as a proof-of-concept and a real world dataset of left atrium MRI scans. In both experiments, we create three different test sets: *control*, *aleatoric*, and *epistemic*. The control test set is well represented under the training population, whereas the epistemic and aleatoric test sets are not. Examples with images that differ from the training images are chosen for the aleatoric set, as this suggests data uncertainty. The epistemic set is chosen to demonstrate model uncertainty by selecting examples with shapes that differ from those in the training set. The test sets are held out from the entire data augmentation and training process. It is important that test sets are not used to build the PDMs, such that they are not reflected in the population statistics captured in the PCA scores. Hence, we use surface-to-surface distances between meshes to quantify shape-based prediction errors since testing samples do not have optimized (ground truth) correspondences.

4.1 Supershapes Dataset

As a proof-of-concept, we construct a set of 3D supershapes shapes [20], which are a family of parameterized shapes. A supershape is parameterized by three variables, one which determines the number of lobes in the shape (or the shapes group), and two which determine the curvature of the shape. To create the training and validation sets, we generate 5000 3-lobe shapes with randomly drawn curvature values (using a χ^2 distribution). For each shape, a corresponding image of size $98 \times 98 \times 98$ is formed, where the intensities of the foreground and background are modeled as Gaussian distributions with different means but same variance. Additive Gaussian noise is added and the images are blurred with a Gaussian filter to mimic diffuse shape boundaries. In this case, Uncertain-DeepSSM predicts a single PCA score, where the first dominant PCA mode captures 99% of the shape variability.

We analyze output uncertainty measures on three different test sets, each of size 100. Examples of these can be seen in Fig. 4. The control test set is generated in the same manner as the training data and provides baseline uncertainty measures. The aleatoric test set contains shapes of the same shape group as the training, but the corresponding images are blurred with a larger Gaussian filter. This makes the shape boundary less clear, which has the effect of adding data uncertainty. For the epistemic test set, the images are blurred to the same degree as the training set, but the shapes belong to a different shape group. Here, we use 5-lobe shapes instead of 3-lobe to demonstrate model uncertainty.

The results of all three test sets are shown in Table 1. The predictions of Uncertain-DeepSSM are more accurate than DeepSSM on all of the test sets, with the aleatoric set having a notable difference. This is a result of the averaging effect of prediction sampling,

which counters the effect of image blurring. The box plots of the uncertainty measure associated with the predicted PCA score in Fig. 5 demonstrate that as expected, Uncertain-DeepSSM predicts higher aleatoric uncertainty on the aleatoric test set and higher epistemic uncertainty on the epistemic test set when compared to the control. The epistemic test has the highest of both forms of uncertainty because changing the shape group produces a great shift in the image domain (aleatoric) and shape domain (epistemic).

4.2 Left Atrium (LA) Dataset

The LA dataset consists of 206 late gadolinium enhancement MRI images of size $235 \times 138 \times 175$ that vary significantly in intensity and quality and have surrounding anatomical structures with similar intensities. The LA shape variation is also significant due to the topological variants pertaining to pulmonary veins arrangements [24]. The variation in images and shapes suggest a strong need for uncertainty measures. For networks training purposes, the images are down-sampled to size $118 \times 69 \times 88$. We predict 19 PCA scores such that 95% of the shape-population variability is preserved. We compare DeepSSM and Uncertain-DeepSSM on three test sets, each of size 30. To define the aleatoric test set, we run PCA (preserving 95% of variability) on all 206 images. We then consider the Mahalanobis distance of the PCA scores of each sample to the mean PCA scores (within-subspace distance) as well as the image reconstruction error (mean square error as off-subspace distance). These values are normalized and summed to get a measure of image similarity to the whole set (similar to [39]). We select the 30 that differ the most to be the aleatoric test set. These examples are the least supported by the input data, suggesting they should have high data uncertainty. To define the epistemic test set, we use the same technique but perform PCA on the signed distance transforms, as an implicit form of shapes, rather than the raw images. In this way, we are able to select an epistemic test set of 30 examples with shapes that are poorly supported by the data. This selection technique produces aleatoric and epistemic test sets that overlap by 6 examples, leaving 152 out of 206 samples. 30 of these are randomly selected to be the control test set and the rest (122) are used in data augmentation to create a training set of 4000 and validation set of 1000. Examples from the test sets can be seen in Fig. 6.

We train both DeepSSM and Uncertain-DeepSSM on different percentages of training data, namely 100%, 75%, and 25%, where an X% is randomly drawn from the remaining 122 samples and then used to proportionally augment the data. The average results of these tests are shown in Table 2. As expected, epistemic uncertainty measures decrease with more training data because model uncertainty can be explained away given more data. Uncertain-DeepSSM made more accurate predictions in most cases, notably when training data is limited. Uncertain-DeepSSM also successfully quantified uncertainty as we can see in the box plots in Fig. 7, which illustrate increased uncertainty measures on the uncertain test sets as compared to the control. The scatter plot in Fig. 7 illustrates the correlation between accuracy and uncertainty measures. The trend lines (combined based on all three test sets) indicate that the uncertainty quantification from Uncertain-DeepSSM provides insight into how trustworthy the model output is.

In Fig. 1, the uncertainty measures are shown on the meshes constructed from model predictions (models trained on 100% of the training data). The top example is from the control set, the second is from the aleatoric set, and the bottom is from the epistemic set. Here, we can see that both aleatoric and epistemic uncertainty are higher in regions where the surface-to-surface distance is higher. This demonstrates the practicality of Uncertain-DeepSSM in a clinical setting as it indicates what regions of the predicted shape professionals can trust and where they should be skeptical.

5 Conclusion

Uncertain-DeepSSM provides a unified framework to predict shape descriptors with measures of both forms of uncertainty directly from 3D images. It maintains the end-to-end nature of DeepSSM while providing an accuracy improvement and uncertainty quantification. By predicting and quantifying uncertainty on PCA scores, Uncertain-DeepSSM enables population-level statistical analysis with aleatoric and epistemic uncertainty measures that can be evaluated in a visually interpretable way. In the future, a layer that maps the PCA scores to the set of correspondence points could be added, enabling fine-tuning the network and potentially providing an accuracy improvement over deterministically mapping PCA scores. Uncertain-DeepSSM bypasses the time-intensive and cost-prohibitive steps of traditional SSM while providing the safety measures necessary to use deep network predictions in clinical settings. Thus, this advancement has the potential to improve medical standards and increase patient accessibility.

Acknowledgement

References

1. Baccetti T, Franchi L, McNamara J: Thin-plate spline analysis of treatment effects of rapid maxillary expansion and face mask therapy in early class iii malocclusions. The European Journal of Orthodontics 21(3), 275–281 (1999) [PubMed: 10407536]

2. Bhalodia R, Dvoracek LA, Ayyash AM, Kavan L, Whitaker R, Goldstein JA: Quantifying the severity of metopic craniosynostosis: A pilot study application of machine learning in craniofacial surgery. Journal of Craniofacial Surgery (2020)

3. Bhalodia R, Elhabian SY, Kavan L, Whitaker RT: Deepssm: A deep learning framework for statistical shape modeling from raw images. CoRR abs/1810.00111 (2018), http://arxiv.org/abs/1810.00111

4. Bhalodia R, Goparaju A, Sodergren T, Whitaker RT, Morris A, Kholmovski E, Marrouche N, Cates J, Elhabian SY: Deep learning for end-to-end atrial fibrillation recurrence estimation. In: Computing in Cardiology, CinC 2018, Maastricht, The Netherlands, September 23–26, 2018 (2018)

5. Bookstein FL: Principal warps: Thin-plate splines and the decomposition of deformations. IEEE Transactions on pattern analysis and machine intelligence 11(6), 567–585 (1989)

6. Bryan R, Nair PB, Taylor M: Use of a statistical model of the whole femur in a large scale, multi-model study of femoral neck fracture risk. Journal of biomechanics 42(13), 2171–2176 (2009) [PubMed: 19646700]

7. Cates J, Bieging E, Morris A, Gardner G, Akoum N, Kholmovski E, Marrouche N, McGann C, MacLeod RS: Computational shape models characterize shape change of the left atrium in atrial fibrillation. Clinical Medicine Insights: Cardiology 8, CMC–S15710 (2014)

8. Cates J, Elhabian S, Whitaker R: Shapeworks: Particle-based shape correspondence and visualization software. In: Statistical Shape and Deformation Analysis, pp. 257–298. Elsevier (2017)

9. Cates J, Fletcher PT, Styner M, Shenton M, Whitaker R: Shape modeling and analysis with entropy-based particle systems. In: IPMI. pp. 333–345. Springer (2007)

10. Chang J, Fisher JW: Efficient mcmc sampling with implicit shape representations. In: CVPR 2011. pp. 2081–2088. IEEE (2011)

11. Davies RH, Twining CJ, Cootes TF, Waterton JC, Taylor CJ: A minimum description length approach to statistical shape modeling. IEEE Transactions on Medical Imaging 21(5), 525–537 (5 2002). 10.1109/TMI.2002.1009388 [PubMed: 12071623]

12. Denker JS, LeCun Y: Transforming neural-net output levels to probability distributions. In: Advances in neural information processing systems. pp. 853–859 (1991)

13. Durrleman S, Prastawa M, Charon N, Korenberg JR, Joshi S, Gerig G, Trouvé A: Morphometry of anatomical shape complexes with dense deformations and sparse parameters. NeuroImage 101, 35–49 (2014) [PubMed: 24973601]

14. Gal Y: Uncertainty in deep learning. University of Cambridge 1, 3 (2016)

15. Gal Y, Ghahramani Z: Bayesian convolutional neural networks with bernoulli approximate variational inference. arXiv preprint arXiv:1506.02158 (2015)

16. Gal Y, Ghahramani Z: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059 (2016)

17. Galloway F, Kahnt M, Ramm H, Worsley P, Zachow S, Nair P, Taylor M: A large scale finite element study of a cementless osseointegrated tibial tray. Journal of biomechanics 46(11), 1900–1906 (2013) [PubMed: 23764172]

18. Gardner G, Morris A, Higuchi K, MacLeod R, Cates J: A point-correspondence approach to describing the distribution of image features on anatomical surfaces, with application to atrial fibrillation. In: 2013 IEEE 10th International Symposium on Biomedical Imaging. pp. 226–229 (4 2013). 10.1109/ISBI.2013.6556453

19. Gerig G, Styner M, Jones D, Weinberger D, Lieberman J: Shape analysis of brain ventricles using spharm. In: Proceedings IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA 2001). pp. 171–178 (2001). 10.1109/MMBIA.2001.991731

20. Gielis J: A generic transformation that unifies a wide range of natural and abstract shapes (9 2013)

21. Glorot X, Bengio Y: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 9, pp. 249–256. PMLR (13–15 5 2010)

22. Harris MD, Datar M, Whitaker RT, Jurrus ER, Peters CL, Anderson AE: Statistical shape modeling of cam femoroacetabular impingement. Journal of Orthopaedic Research 31(10), 1620–1626 (2013). 10.1002/jor.22389, 10.1002/jor.22389 [PubMed: 23832798]

23. He K, Zhang X, Ren S, Sun J: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. CoRR abs/1502.01852 (2015), http://arxiv.org/abs/1502.01852

24. Ho SY, Cabrera JA, Sanchez-Quintana D: Left atrial anatomy revisited. Circulation: Arrhythmia and Electrophysiology 5(1), 220–228 (2012) [PubMed: 22334429]

25. Huang W, Bridge CP, Noble JA, Zisserman A: Temporal heartnet: Towards human-level automatic analysis of fetal cardiac screening video. In: MICCAI 2017. pp. 341–349. Springer International Publishing (2017)

26. Joshi SC, Miller MI: Landmark matching via large deformation diffeomorphisms. IEEE Transactions On Image Processing 9(8), 1357–1370 (8 2000) [PubMed: 18262973]

27. Joshi S, Davis B, Jomier M, Gerig G: Unbiased diffeomorphic atlas construction for computational anatomy. NeuroImage 23(Supplement1), S151–S160 (2004) [PubMed: 15501084]

28. Kendall A, Gal Y: What uncertainties do we need in bayesian deep learning for computer vision? CoRR abs/1703.04977 (2017), http://arxiv.org/abs/1703.04977

29. Kingma D, Ba J: Adam: A method for stochastic optimization. International Conference on Learning Representations (12 2014)

30. Kiureghian AD, Ditlevsen OD: Aleatory or epistemic? does it matter? (2009)

31. Kozic N, Weber S, Büchler P, Lutz C, Reimers N, Ballester MÁG, Reyes M: Optimisation of orthopaedic implant design using statistical shape space analysis based on level sets. Medical image analysis 14(3), 265–275 (2010) [PubMed: 20359938]

32. Kwon Y, Won JH, Kim BJ, Paik MC: Uncertainty quantification using bayesian neural networks in classification: Application to ischemic stroke lesion segmentation (2018)

33. Lamecker H, Lange T, Seebass M: A statistical shape model for the liver. In: International conference on medical image computing and computer-assisted intervention. pp. 421–427. Springer (2002)

34. Lê M, Unkelbach J, Ayache N, Delingette H: Sampling image segmentations for uncertainty quantification. Medical image analysis 34, 42–51 (2016) [PubMed: 27198913]

35. Le QV, Smola AJ, Canu S: Heteroscedastic gaussian process regression. In: Proceedings of the 22nd international conference on Machine learning. pp. 489–496 (2005)

36. Li X, Chen S, Hu X, Yang J: Understanding the disharmony between dropout and batch normalization by variance shift. CoRR abs/1801.05134 (2018), http://arxiv.org/abs/1801.05134

37. MacKay DJ: A practical bayesian framework for backpropagation networks. Neural computation 4(3), 448–472 (1992)

38. Milletari F, Rothberg A, Jia J, Sofka M: Integrating statistical prior knowledge into convolutional neural networks. In: Descoteaux M, Maier-Hein L, Franz A, Jannin P, Collins DL, Duchesne S (eds.) Medical Image Computing and Computer Assisted Intervention. pp. 161–168. Springer International Publishing, Cham (2017)

39. Moghaddam B, Pentland A: Probabilistic visual learning for object representation. IEEE Transactions on pattern analysis and machine intelligence 19(7), 696–710 (1997)

40. Nix DA, Weigend AS: Estimating the mean and variance of the target probability distribution. In: Proceedings of 1994 ieee international conference on neural networks (ICNN'94). vol. 1, pp. 55–60. IEEE (1994)

41. Oktay O, Ferrante E, Kamnitsas K, Heinrich MP, Bai W, Caballero J, Guerrero R, Cook SA, de Marvao A, Dawes T, O'Regan DP, Kainz B, Glocker B, Rueckert D: Anatomically constrained neural networks (ACNN): application to cardiac image enhancement and segmentation. CoRR abs/1705.08302 (2017), http://arxiv.org/abs/1705.08302

42. Reinhold JC, He Y, Han S, Chen Y, Gao D, Lee J, Prince JL, Carass A: Validating uncertainty in medical image translation. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). pp. 95–98. IEEE (2020)

43. Sarkalkan N, Weinans H, Zadpoor AA: Statistical shape and appearance models of bones. Bone 60, 129–140 (2014) [PubMed: 24334169]

44. Selvan R, Faye F, Middleton J, Pai A: Uncertainty quantification in medical image segmentation with normalizing flows. arXiv preprint arXiv:2006.02683 (2020)

45. Styner M, Oguz I, Xu S, Brechbuehler C, Pantazis D, Levitt J, Shenton M, Gerig G: Framework for the statistical shape analysis of brain structures using spharm-pdm (07 2006)

46. Tóthová K, Parisot S, Lee MCH, Puyol-Antón E, Koch LM, King AP, Konukoglu E, Pollefeys M: Uncertainty quantification in cnn-based surface prediction using shape priors. CoRR abs/1807.11272 (2018), http://arxiv.org/abs/1807.11272

47. Wang D, Shi L, Griffith JF, Qin L, Yew DT, Riggs CM: Comprehensive surface-based morphometry reveals the association of fracture risk and bone geometry. Journal of Orthopaedic Research 30(8), 1277–1284 (2012) [PubMed: 22253193]

48. Xie J, Dai G, Zhu F, Wong EK, Fang Y: Deepshape: deep-learned shape descriptor for 3d shape retrieval. IEEE transactions on pattern analysis and machine intelligence 39(7), 1335–1345 (2017) [PubMed: 27482929]

49. Zhao Z, Taylor WD, Styner M, Steffens DC, Krishnan KRR, MacFall JR: Hippocampus shape analysis and late-life depression. PLoS One 3(3) (2008)
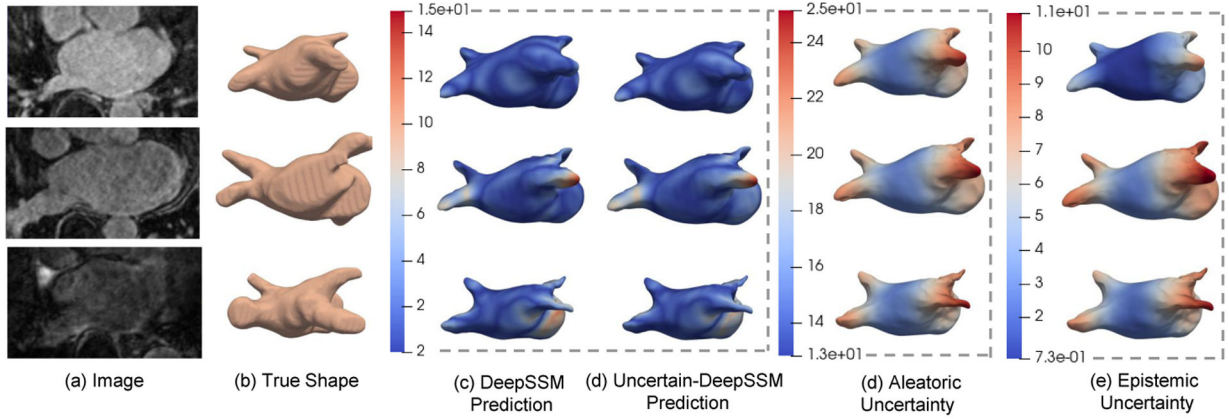
50. Zheng Y, Liu D, Georgescu B, Nguyen H, Comaniciu D: 3d deep learning for efficient and robust landmark detection in volumetric data. In: MICCAI 2015. pp. 565–572. Springer International Publishing (2015)

(a) Image (b) True Shape (c) DeepSSM Prediction (d) Uncertain-DeepSSM Prediction (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

Fig.1:

Shape prediction and uncertainty quantification on left atrium MRI scans. The images (a) are input and the true shapes (surface meshes) (b) are from ground truth segmentations. Shapes in (c) and (d) are constructed from DeepSSM and Uncertain-DeepSSM predictions, respectively. The heat maps on surface meshes in (c) and (d) show the surface-to-surface distance to (b) (the error in mm). The aleatoric (d) and epistemic (e) output from Uncertain-DeepSSM are shown as heat maps on the predicted mesh. Our model outputs increased uncertainty where error is high.
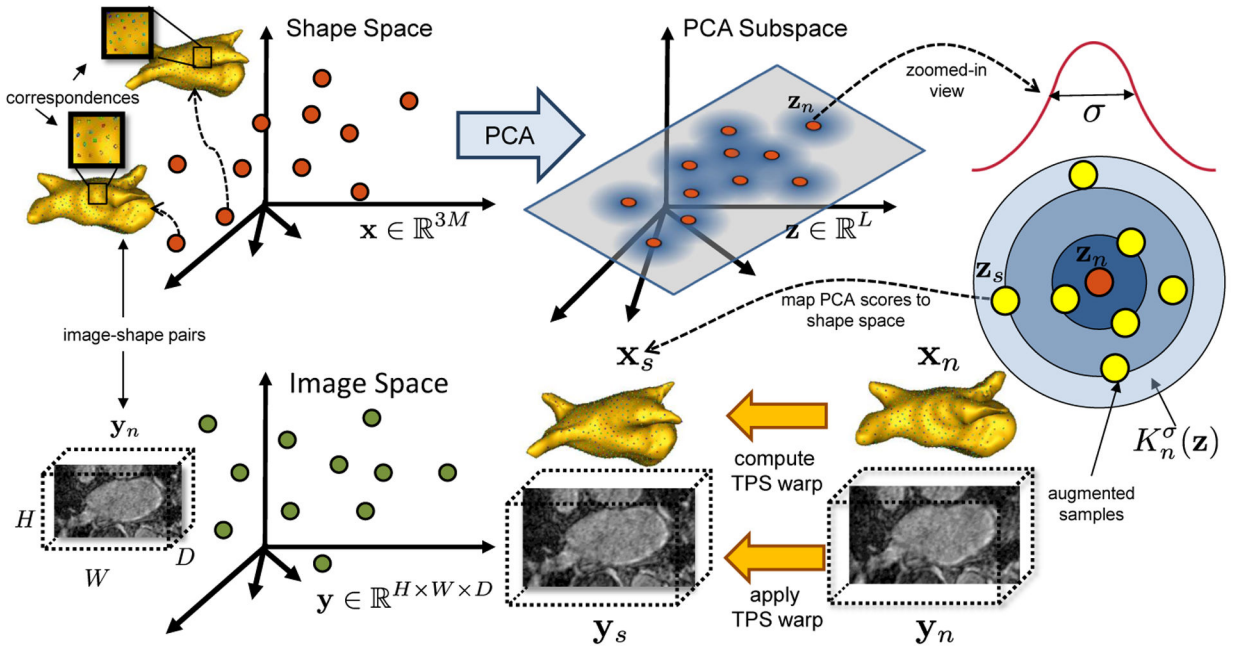
Fig.2:

Data augmentation: PCA is used to compute the PCA scores $\{\mathbf{z}_n\}_{n=1}^{N}$ of the training shape samples $\{\mathbf{x}_n\}_{n=1}^{N}$. Augmented samples $\mathbf{z}_s$ are randomly drawn from a collection of multivariate Gaussian distributions $K_\sigma^n(\mathbf{z})$, each with a training example $\mathbf{z}_n$ as the mean and covariance $\sigma^2 \mathbf{I}_L$. The correspondences $\mathbf{x}_s$ are used to compute a TPS warp to map $\mathbf{x}_n$ to $\mathbf{x}_s$, which is used to warp the respective image $\mathbf{y}_n$ to a new image $\mathbf{y}_s$ with known shape parameters $\mathbf{z}_s$.
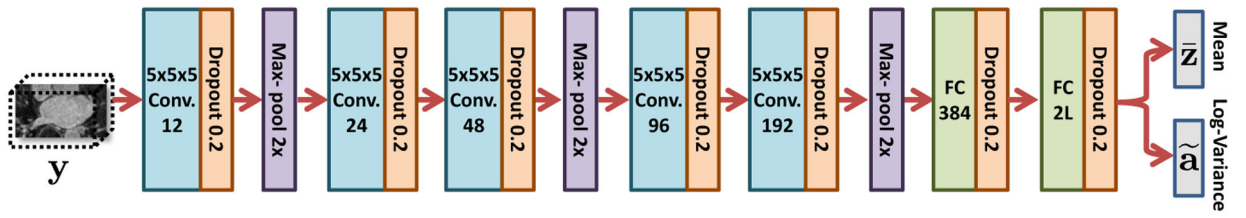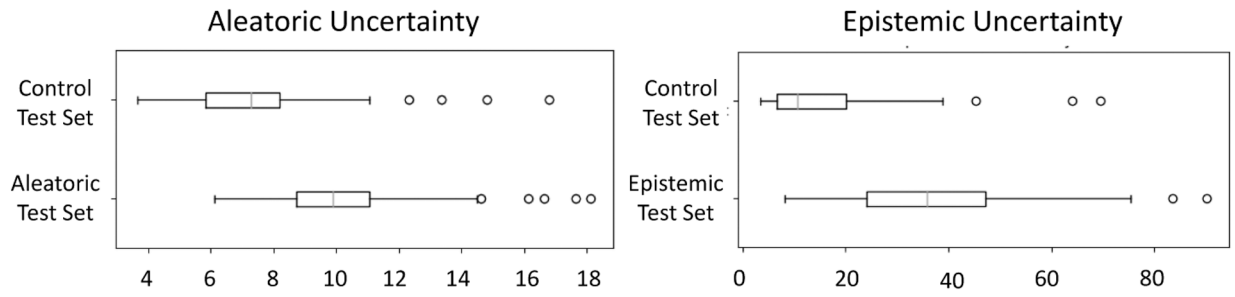
Fig.3:
Uncertain-DeepSSM network architecture
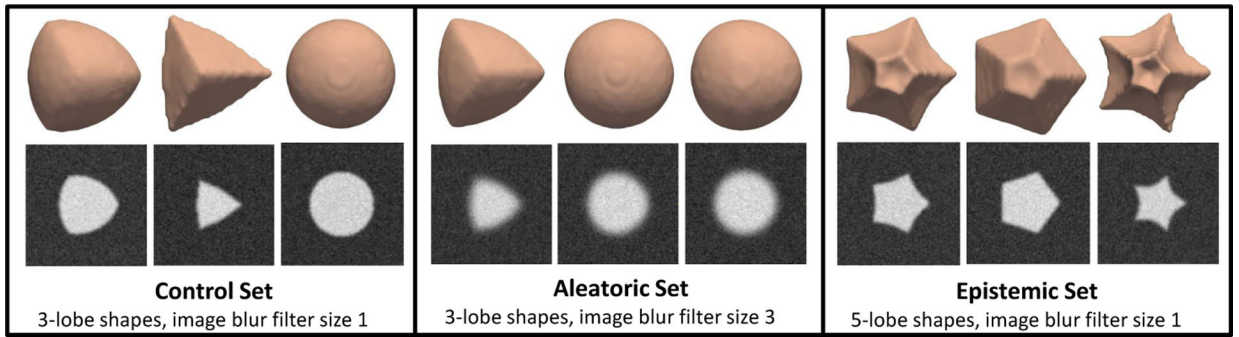
Fig.4:
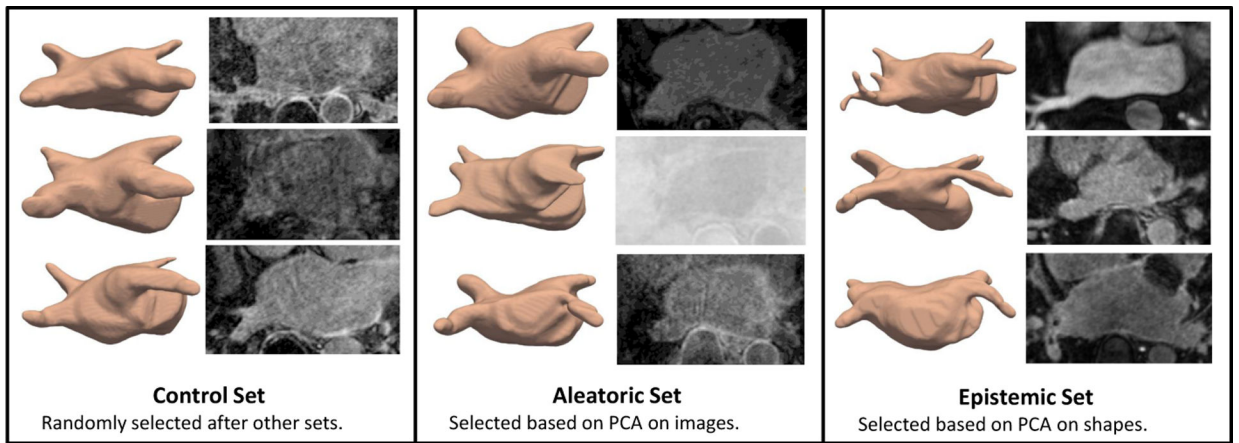Examples of surfaces and corresponding image slices from supershapes test sets.

Fig.5:
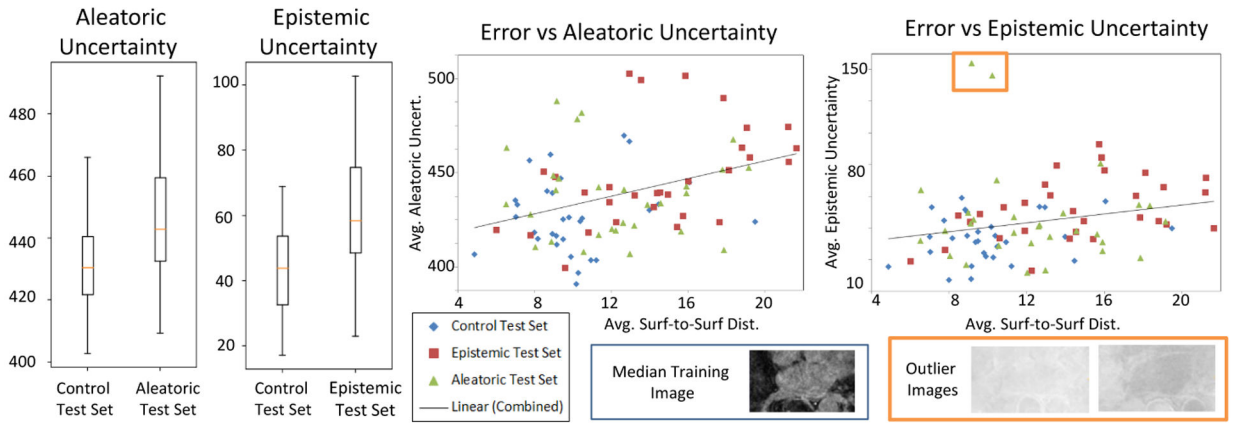Boxplots of supershapes uncertainties compared to control test set.

Fig.6:

Examples from of surfaces and image slices for LA test sets.

Fig.7:

Results on LA test sets from training Uncertain-DeepSSM on the full training set. The box plots show output uncertainty measures compared to the control test set. The scatter plot shows the average error versus average uncertainty on all three test sets. The two outliers marked with an orange box in the epistemic uncertainty plot are examples with images of a much higher intensity than the training examples (shown to the right) causing a spike in epistemic uncertainty.

Table 1:

Average error and uncertainty measures on supershapes test sets.

	DeepSSM	Uncertain-DeepSSM		
	Surface-to-Surface Distance	Surface-to-Surface Distance	Aleatoric Uncertainty	Epistemic Uncertainty
Control Test Set	0.670 ± 0.104	0.615 ± 0.163	7.413 ± 2.189	15.000 ± 11.762
Aleatoric Test Set	1.293 ± 0.679	0.798 ± 0.447	10.205 ± 2.276	22.178 ± 13.065
Epistemic Test Set	7.045 ± 1.653	7.008 ± 1.668	12.256 ± 5.424	36.226 ± 17.327

Table 2:

Results on left atrium test sets with various training set sizes. Reported surface-to-surface distances are averaged across the test set and uncertainty measures are averaged across PCA modes and the test set.

		DeepSSM	Uncertain-DeepSSM		
		Surface-to-Surface Distance (mm)	Surface-to-Surface Distance (mm)	Aleatoric Uncertainty	Epistemic Uncertainty
Control Test Set	25% Train	15.262 ± 3.694	10.670 ± 2.560	519.026 ± 7.357	58.206 ± 38.145
	75% Train	10.319 ± 2.834	10.072 ± 2.812	452.025 ± 0.519	46.581 ± 32.678
	100% Train	10.205 ± 2.779	10.153 ± 2.904	431.518 ± 0.674	43.561 ± 29.821
Aleatoric Test Set	25% Train	12.967 ± 3.592	12.830 ± 3.543	472.359 ± 10.164	64.312 ± 45.656
	75% Train	12.507 ± 3.522	12.169 ± 3.493	465.951 ± 1.089	60.458 ± 45.454
	100% Train	12.242 ± 3.602	12.289 ± 3.525	442.129 ± 0.917	56.009 ± 42.371
Epistemic Test Set	25% Train	15.759 ± 4.301	14.975 ± 4.209	465.817 ± 7.567	75.854 ± 51.188
	75% Train	14.690 ± 4.166	14.581 ± 4.104	446.641 ± 1.127	64.236 ± 44.642
	100% Train	14.558 ± 4.151	14.465 ± 4.092	448.082 ± 1.291	61.517 ± 42.259