Weakly SSM : On the Viability of Weakly Supervised Segmentations for Statistical Shape Modeling

Janmesh Ukey, Tushar Kataria, and Shireen Y. Elhabian[3]

[1] Kahlert School of Computing, University of Utah, Salt Lake City, USA
[2] Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, USA
[3] Corresponding author
{janmesh,tushar.kataria,shireen}@sci.utah.edu

Abstract. Statistical Shape Models (SSMs) excel at identifying population level anatomical variations, which is at the core of various clinical and biomedical applications, including morphology-based diagnostics and surgical planning. However, the effectiveness of SSM is often constrained by the necessity for expert-driven manual segmentation, a process that is both time-intensive and expensive, thereby restricting their broader application and utility. Recent deep learning approaches enable the direct estimation of Statistical Shape Models (SSMs) from unsegmented images. While these models can predict SSMs without segmentation during deployment, they do not address the challenge of acquiring the manual annotations needed for training, particularly in resource-limited settings. Semi-supervised and foundation models for anatomy segmentation can mitigate the annotation burden. Yet, despite the abundance of available approaches, there are no established guidelines to inform end-users on their effectiveness for the downstream task of constructing SSMs. In this study, we systematically evaluate the potential of weakly supervised methods as viable alternatives to manual segmentation's for building SSMs. We establish a new performance benchmark by employing various semi-supervised and foundational model methods for anatomy segmentation under low annotation settings, utilizing the predicted segmentation's for the task of SSM. We compare the modes of shape variation and use quantitative metrics to compare against a shape model derived from a manually annotated dataset. Our results indicate that some methods produce noisy segmentation, which is very unfavorable for SSM tasks, while others can capture the correct modes of variations in the population cohort with 60-80% reduction in required manual annotation.

Keywords: Statistical Shape Modeling · Semi Supervised Segmentation

1 Introduction

Statistical shape models (SSMs) provide a quantitative means for analyzing anatomical variations across populations, enabling the identification of norma-

tive trends and deviations and facilitating the development of diagnostic tools and surgical planning systems [10]. Constructing SSMs is contingent upon the accurate segmentation of the target anatomy. This process is both time-consuming and resource-intensive, often hindered by the scarcity of medical expertise necessary for precise segmentation. Recent advances in deep learning have facilitated the direct estimation of SSMs from unsegmented images, thus bypassing the need for segmentation during inference [7,2,3,21,19,28,20,13,6,23,24,6,29,12,11,4]. However, these deep learning methods still necessitate anatomy segmentation to construct SSMs for training.

Automated or deep learning based anatomy segmentation can mitigate the segmentation burden for constructing SSMs; however, networks designed for segmentation task still require significant manual annotations for their training. To alleviate such burden, a variety of semi-supervised approaches have been developed, each demonstrating different levels of performance. *Semi-supervised methods* (e.g., [30,15,5,31,22,32,25,27,18]), typically leverage a subset of fully annotated volumes alongside all unannotated volumes for model training. These methods use a combination of clever data augmentations [5], pseudo labelling[5,22,31], consistency regularization[22,27], and entropy minimization[25] to obtain high quality segmentation's. A few weakly supervised methods (e.g., [15]) exploit the inherent 3D structure of the input image, allowing for the reduction of annotation requirements from entire volumes to individual slices.

As deep learning models scale up to billions of parameters trained on millions of data samples annually, foundational models emerge for optimal generalization across unseen datasets. SAM [14], for instance, accepts bounding box or point prompts to predict anatomy segmentation and has been fine-tuned on medical data, effectively reducing annotation requirements.Currently, numerous annotation-efficient methods exist for anatomy segmentation. However, there are no established guidelines to evaluate their suitability as alternatives to manual segmentation for constructing Statistical Shape Models (SSMs). Accurate delineation of the target anatomy is crucial for revealing population-level statistics in SSM construction.
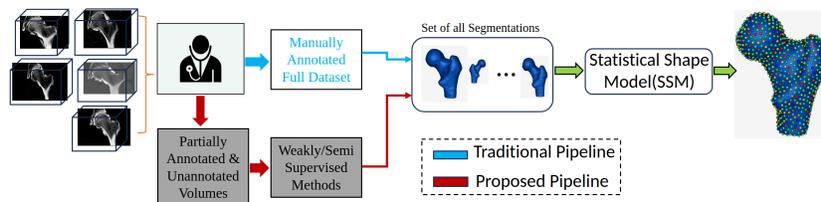

Fig. 1: **Can manual annotation in constructing SSM be replaced by weakly supervised segmentation methods?** Traditional SSM pipelines necessitate manual annotations by medical professionals on the entire dataset. We propose leveraging partially annotated and unannotated volumes to train weakly supervised segmentation models, subsequently use these models to produce 3D anatomy segmentations to construct the shape models.

In this paper, we introduce a comprehensive performance benchmark for constructing SSMs using annotations predicted using weakly/semi-supervised segmentation methods. By adopting weak supervision techniques such as semi-supervised methods we aim to alleviate the annotation requirements for constructing SSMs, thus increasing the applicability and usability of SSM tools. Our analysis seeks to identify which weakly supervised approaches are viable for deployment in population-level analysis tasks. Our benchmark evaluates a variety of techniques, including semi-supervised methods and foundational models, against Statistical Shape Models (SSMs) constructed from manual segmentations. This assessment aims to gauge their effectiveness and feasibility in scenarios with limited annotations. Our analysis reveals that certain methods yield poor segmentations due to sensitivity to hyperparameter settings and biases for image modality. Conversely, other approaches closely approximate SSM models trained on complete annotated datasets, successfully capturing the population cohort's modes of variation, albeit with a degree of noise. The main contributions of this paper are:

- We establish a new benchmark for Semi-supervised SSM by comparing SSMs trained with manual segmentations against those derived from semi-supervised methods across two datasets.
- We evaluate SSM models trained with varying quantities and types of manual segmentations 20% or 40% of manual full-volume annotations for semi-supervised models.
- We assess and identify which methods can reliably replace manual annotations in scenarios with limited resources.

2   Weak Supervision Methods

To investigate whether weakly supervised methods can be safely substituted instead of manual annotations for SSM, we address the following questions :-

- *Do SSM models created using weak supervision predictions capture the same modes of variation compared to an SSM model using all manual segmentations ? Are some weakly supervised methods more reliable than others ?*
- *Does increasing the amount of annotated data for weakly supervised models improve SSM for the population cohort?*
- *Do quantitative metrics dice correlate better with better accuracy on downsteam task of SSM ?*

This section gives a brief summary of semi-supervised methods used in the study.

2.1   Semi-Supervised Methods

Semi-supervised medical image segmentation involves utilization of partially labeled and unlabeled data in the training process. Majority of semi-supervised methods are based on *student-teacher* paradigm. These methods employ self-training by pseudo labelling or consistency regularization/co-training by entropy

minimization of predictions using multiple views of the input images. We picked seven unique methods that use labeled and unlabelled data in different ways.

Mean Teacher (MT)[22]. Mean Teacher uses consistency regularization between *student*-**teacher** training paradigm. The *student* model in trained using the supervised loss on the labelled set, where as the *teacher*'s parameters are updated as an exponential moving average of the student model. Consistency loss on unlabelled data aligns the prediction between the two models.

Uncertainity Mean Teacher (MT-UC)[31]. An extension of MT where the *teacher* model also estimates the uncertainty of each target prediction with Monte Carlo sampling. The uncertainty is used to preserve only the reliable predictions when calculating the consistency loss.

Bidirectional Copy-Paste (BCP)[5]. BCP integrates a bidirectional copy-paste framework into the Mean Teacher architecture. The student network receives inputs created by pasting random crops from labeled images onto unlabeled images and vice versa (bi-direction copy). The supervision of the *student* network involves combining both ground-truth labels and pseudo-labels generated by the *teacher* network using the same bidirectional copy-paste process.

Deep Adversarial Networks (AN) [32]. AN comprises two networks: a Segmentation Network (SN) for performing segmentation and an Evaluation Network (EN) for assessing segmentation quality. During training, EN learns to distinguish between segmentations of labeled and unlabeled images, while SN is trained to produce segmentations for unlabeled images that EN cannot differentiate from those of labeled images. This iterative adversarial training process, where EN critiques segmentations of unlabeled images, trains SN to generate more accurate segmentations for both unlabeled and unseen samples, resulting in improved segmentation performance.

Entropy Minimization (EM) [25]. In semi-supervised segmentation, entropy minimization involves using entropy loss to penalize low-confidence predictions on unlabeled samples. The entropy loss is calculated from the model's pixel-wise predictions, reflecting the uncertainty or confidence in its segmentation outputs. By minimizing the entropy loss, the model is encouraged to make more certain and accurate predictions, especially on unlabeled samples where it may initially show higher uncertainty.

Regularized Dropout (RD) [27]. RD consists of two networks with different initial weight configurations. During training, both models are supervised with a segmentation loss between their predictions and the ground truth for labeled samples. For unlabeled data, a consistency loss using KL divergence measures the agreement between the models' outputs.

Uncertainty Rectified Pyramid Consistency (URPC) [18]. URPC utilizes a network that provides multi-scale predictions and enforces consistency among these predictions for both labeled and unlabeled data. A standard supervised loss is employed for learning from labeled images, while for unlabeled images, the model is encouraged to produce consistent multi-scale predictions as a regularization strategy. An Uncertainty Rectified Pyramid Consistency (URPC) module is incorporated to enhance model stability and performance by address-

ing unreliable predictions in unlabeled data. The uncertainty estimation is based on measuring the discrepancy between the model's predictions at different scales, allowing for the selection of reliable voxels for loss calculation and ensuring stable unsupervised training.

SAM Med3D (SAM) [26]. SAM-Med3D transforms SAM's original 2D components into 3D counterparts, including a 3D image encoder, a 3D prompt encoder, and a 3D mask decoder. The 3D image encoder uses 3D convolutions with a learnable 3D absolute positional encoding, extending SAM's 2D positional encoding to capture spatial information. The 3D prompt encoder handles sparse prompts with 3D positional encodings and dense prompts with 3D convolutions. The 3D mask decoder employs 3D upscaling procedures and 3D transposed convolutions. The SAM-Med3D model was trained from scratch using large-scale volumetric medical data.

VNet Baseline. [1] VNet is a fully supervised method for volumetric medical image segmentation. It has a V-shaped architecture with residual connections in each layer. It consists of a compression path on the left side and a decompression path on the right side. The network operates in different stages, each comprising one to three convolutional layers. Each stage learns a residual function by adding the input to the output of the last convolutional layer. The convolutions in each stage use volumetric kernels of size 5x5x5 voxels, and the resolution of the data is reduced as it progresses through the compression path.

3   Results and Discussion

Dataset Details. Two dataset were used for analysis, one public and one in-house dataset. We used the NAMIC public Left Atrium segmentation dataset used in [6]. Our in-house dataset consisted of 49 volumes, 40 for train and rest 9 for test. We experiment two different percent of labeled samples, more specifically we utilize 20% (8 labelled and 32 unlabelled) and 40% (16 labelled and 24 unlabelled) for training the semi-supervised models. For the public Left Atrium dataset, we divide into 50 train and 9 for test. Similarly we experiment two different percent of labeled samples, more specifically we utilize 20% (10 labelled and 40 unlabelled) and 40% (20 labelled and 30 unlabelled) for training the semi-supervised models.

Implementation Details. Original implementation with default parameters were used for Mean Teacher, Uncertainty Aware Mean Teacher, Bidirectional Copy-Paste, Uncertainty Rectified Pyramid Consistency and SAM-Med3D. For Deep Adversarial Networks, Entropy Minimization and Regularized Dropout implementation provided by [17] with default parameters were used. For SAM-Med3D 10 random points in foreground region were provided for inference. ShapeWorks [8] was choosen as the tool to create all shape models for analysis because of it's high efficacy [10]. The shape models were formed using the prediction of unlabeled samples from training set.

Metric Used We employ the quantitative evaluation framework proposed by [9] to assess shape models. This framework entails the examination of various met-

rics including compactness, generalization, and specificity. Generalization and specificity distance is reported in mm. These distance calculated as distance between SSM correspondences to closest surface point of the ground truth mesh. These metrics, as a function of modes of variation, are derived from principal component analysis (PCA) conducted on shape correspondences. Furthermore, we expand our evaluation methodology by integrating a measure of subspace distance. Specifically, we utilize the concept of Grassmannian distance as introduced by [16]. The Grassmannian distance are calculated for each method from the PCA space constructed from the ground truth segmentation's. For qualitative evaluation we utilize modes of variation to examine the shape variability capture by the first two dominant modes.

3.1   Results and Discussion

Segmentation Results: We report the Dice score on unlabeled data samples in Table 1 for both 20% and 40% of labeled samples used in training the semi-supervised method. It is evident that BCP is the best-performing semi-segmentation methods when using 20% of labeled samples for both the femur and left atrium. However, we observe a significant drop in BCP's performance for the left atrium as the number of labeled samples increases. This decline in performance with an increasing number of labeled samples is also noted in other semi-supervised methods. V-Net, trained from scratch using only labeled samples, outperforms some of the semi-supervised methods. This raises questions about the efficacy of these semi-supervised approaches.

Table 1: **Semi-Supervised Segmentation Results.** Dice Score on the unlabeled samples reported for all methods used in the study.

Dataset	% Data	BCP	RD	AN	EM	MT	MT-UC	URPC	SAM10	V-Net
Femur	20 (8 labeled)	0.959	0.923	0.920	0.923	0.917	0.909	0.916	0.889	0.921
	40 (16 labeled)	0.965	0.934	0.898	0.906	0.919	0.924	0.954	-	0.918
Left Atrium	20 (10 labeled)	0.903	0.885	0.865	0.884	0.879	0.873	0.878	0.825	0.874
	40 (20 labeled)	0.683	0.886	0.879	0.895	0.874	0.882	0.901	-	0.881

Statistical Shape modelling Results. SSMs were created using predicted segmentations from various semi-supervised methods and compared with SSMs derived from manual annotations of the full dataset. Qualitative results for first and second mode of variation for Femur and Left Atrium are shown in Figure 2 and Supplementary Figure 6. Qualitative results indicate that predictions from certain semi-supervised methods (MT, MT-UC, URPC, EM, and AN) produce non-smooth surfaces. This affects the statistical shape model (SSM), causing the first mode of variation captured by the SSM to reflect prediction noise rather than diagnostically relevant information. We observe that while VNet and RD results are smoother, the modes of variation they capture differ from the ground truth SSM(obtained through manual segmentation). The BCP model

most closely captures the mode of variation for the femur with SAM being the second best model. However, both SAM and BCP fails to do so for the pulmonary vein in the left atrium. This discrepancy is likely due to the pulmonary vein being a small protuberance in the dataset, leading to poor segmentation and causing all models to overlook the variations in the dataset. Qualitative results for generalization and specificity are shown in Supplementary Figure 5.
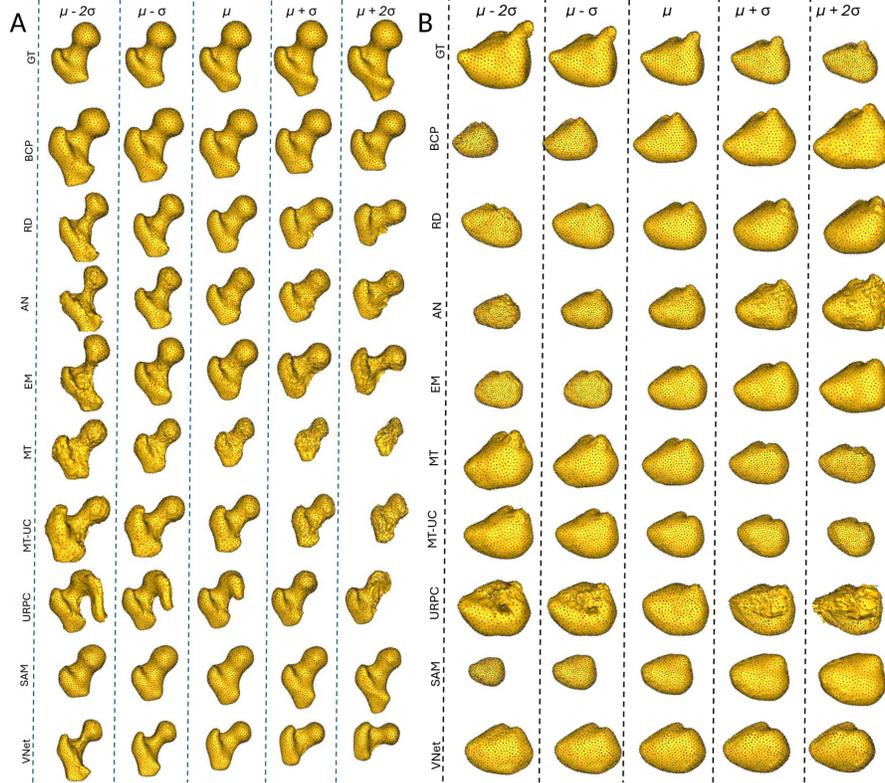

Fig. 2: **First Mode Of variation Results for Femur and Left Atrium using 20% of labelled data.** We show first mode of variations for both Femur and Left Atrium datasets showing mean shape(μ), first ($\pm\sigma$) and second order ($\pm2\sigma$) variations from mean shape.

Different metrics quantifying compactness, generalization, specificity and grassmanian distance for femur and left atrium are shown in Figure 3 and 4. All metrics are calculated with different modes of variations found using PCA decomposition. BCP is the closest model to manual segmentation SSM across different metrics for all modes of variations. However, SAM is second best in specificity and regularized dropout is second best in generalization. There is no clear pattern across all metrics. Grassmanian distance doesn't show any model being hugely different from ground truth.

Correlation between Dice and SSM quantitative metrics. For the femur, lower specificity and a shorter generalization distance are correlated with a better Dice score, although this relationship does not apply to compactness or Grassmannian distance. Instead, it depends on factors such as the size of the labeled dataset and the predicted segmentation quality of the method, particularly in capturing the correct finer details of the underlying segmented anatomy.
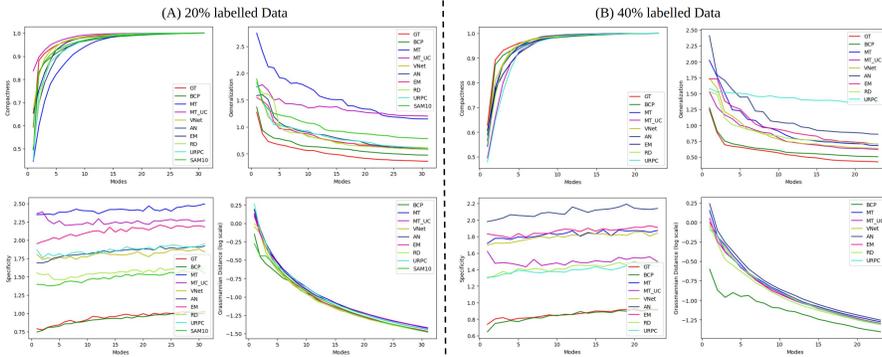

Fig. 3: **Femur Results**. Compactness, Generalization Distance, Specificity Distance and Grassmanian results using (A) 20% of labelled samples (B) 40% of labelled samples for training different Semi-supervised methods.

4   Conclusion

In this study, we address an important bottleneck limiting wide acceptance of Statistical shape modelling for diagnostic decisions, i.e., manual annotation burden. We offer semi-supervised and foundational models as alternatives to manual segmentation and evaluate several of these models on two datasets. Through qualitative and quantitative comparisons, our results demonstrated that some of these weakly supervised methods can indeed serve as viable alternatives for manual segmentation, thus reducing the annotation burden. Notably, our evaluation highlighted the effectiveness of BCP [5] and SAM [26] as a particularly promising methods capturing the correct mode of variations. This finding highlights BCP's [5] potential as a dependable alternative to manual segmentation for certain anatomies. However, BCP struggles to capture intricate details in the complex anatomy of the left atrium, leaving room for future improvements. We intend to broaden this comparison by incorporating a wider array of weakly supervised segmentation approaches, including self-supervised, unsupervised methods, transfer learning, and other foundational models. By exploring this diverse spectrum of segmentation techniques, we aim to gain deeper insights into their effectiveness and applicability for downstream diagnostics tasks such as statistical shape modeling.

References

1. Abdollahi, A., Pradhan, B., Alamri, A.: Vnet: An end-to-end fully convolutional neural network for road extraction from high-resolution remote sensing data. Ieee Access **8**, 179424–179436 (2020)
2. Adams, J., Bhalodia, R., Elhabian, S.: Uncertain-deepssm: From images to probabilistic shape models. In: Shape in Medical Imaging: International Workshop, ShapeMI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings. pp. 57–72. Springer (2020)
3. Adams, J., Elhabian, S.: From images to probabilistic anatomical shapes: A deep variational bottleneck approach. arXiv preprint arXiv:2205.06862 (2022)
4. Adams, J., Iyer, K., Elhabian, S.: Weakly supervised bayesian shape modeling from unsegmented medical images. arXiv preprint arXiv:2405.09697 (2024)
5. Bai, Y., Chen, D., Li, Q., Shen, W., Wang, Y.: Bidirectional copy-paste for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11514–11524 (2023)
6. Bhalodia, R., Elhabian, S., Adams, J., Tao, W., Kavan, L., Whitaker, R.: Deepssm: A blueprint for image-to-shape deep learning models. Medical Image Analysis **91**, 103034 (2024)
7. Bhalodia, R., Elhabian, S.Y., Kavan, L., Whitaker, R.T.: Deepssm: a deep learning framework for statistical shape modeling from raw images. In: Shape in Medical Imaging: International Workshop, ShapeMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings. pp. 244–257. Springer (2018)
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. Davies, R.H., Twining, C.J., Cootes, T.F., Waterton, J.C., Taylor, C.J.: A minimum description length approach to statistical shape modeling. IEEE Trans. Med. Imaging **21**(5), 525–537 (May 2002). https://doi.org/10.1109/TMI.2002.1009388
10. Goparaju, A., Iyer, K., Bone, A., Hu, N., Henninger, H.B., Anderson, A.E., Durrleman, S., Jacxsens, M., Morris, A., Csecs, I., et al.: Benchmarking off-the-shelf statistical shape modeling tools in clinical applications. Medical Image Analysis **76**, 102271 (2022)
11. Iyer, K., Adams, J., Elhabian, S.Y.: Scorp: Statistics-informed dense correspondence prediction directly from unsegmented medical images. arXiv preprint arXiv:2404.17967 (2024)
12. Iyer, K., Elhabian, S.Y.: Mesh2ssm: From surface meshes to statistical shape models of anatomy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 615–625. Springer (2023)
13. Karanam, M.S.T., Kataria, T., Iyer, K., Elhabian, S.Y.: Adassm: Adversarial data augmentation in statistical shape models from images. In: International Workshop on Shape in Medical Imaging. pp. 90–104. Springer (2023)
14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
15. Li, S., Cai, H., Qi, L., Yu, Q., Shi, Y., Gao, Y.: Pln: Parasitic-like network for barely supervised medical image segmentation. IEEE Transactions on Medical Imaging **42**(3), 582–593 (2022)

16. Lim, L.H., Wong, K.S.W., Ye, K.: The grassmannian of affine subspaces. Found. Comput. Math. **21**, 537–574 (Mar 2021). https://doi.org/10.1007/s10208-020-09459-8
17. Luo, X.: SSL4MIS. `https://github.com/HiLab-git/SSL4MIS` (2020)
18. Luo, X., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Chen, N., Wang, G., Zhang, S.: Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. pp. 318–329 (2021)
19. Milletari, F., Rothberg, A., Jia, J., Sofka, M.: Integrating statistical prior knowledge into convolutional neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 161–168. Springer (2017)
20. Raju, A., Miao, S., Jin, D., Lu, L., Huang, J., Harrison, A.P.: Deep implicit statistical shape models for 3d medical image delineation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2135–2143 (2022)
21. Tao, W., Bhalodia, R., Elhabian, S.: Learning population-level shape statistics and anatomy segmentation from images: A joint deep learning model. arXiv preprint arXiv:2201.03481 (2022)
22. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
23. Ukey, J., Elhabian, S.: Localization-aware deep learning framework for statistical shape modeling directly from images. In: Medical Imaging with Deep Learning (2023)
24. Ukey, J., Kataria, T., Elhabian, S.Y.: MASSM: An end-to-end deep learning framework for multi-anatomy statistical shape modeling directly from images (2024). https://doi.org/10.48550/arXiv.2403.11008, retrieved from URL
25. Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P.: Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2517–2526 (2019)
26. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., Fu, B., Zhang, S., He, J., Qiao, Y.: Sam-med3d (2023)
27. Wu, L., Li, J., Wang, Y., Meng, Q., Qin, T., Chen, W., Zhang, M., Liu, T.Y., et al.: R-drop: Regularized dropout for neural networks. Advances in Neural Information Processing Systems **34**, 10890–10905 (2021)
28. Xie, J., Dai, G., Zhu, F., Wong, E.K., Fang, Y.: Deepshape: Deep-learned shape descriptor for 3d shape retrieval. IEEE transactions on pattern analysis and machine intelligence **39**(7), 1335–1345 (2016)
29. Xu, H., Elhabian, S.Y.: Image2ssm: Reimagining statistical shape models from images with radial basis functions. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 508–517. Springer (2023)
30. Xu, Z., Niethammer, M.: Deepatlas: Joint semi-supervised learning of image registration and segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22. pp. 420–429. Springer (2019)
31. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Confer-

ence, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22. pp. 605–613. Springer (2019)

32. Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z.: Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In: Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20. pp. 408–416. Springer (2017)
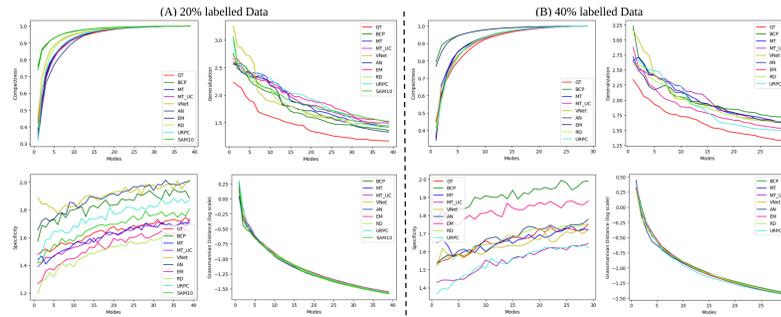

Fig. 4: **Left Atrium Results** Compactness, Generalization, Specificity and Grassmanian results using (A) 20% of labelled samples (B) 40% of labelled samples.
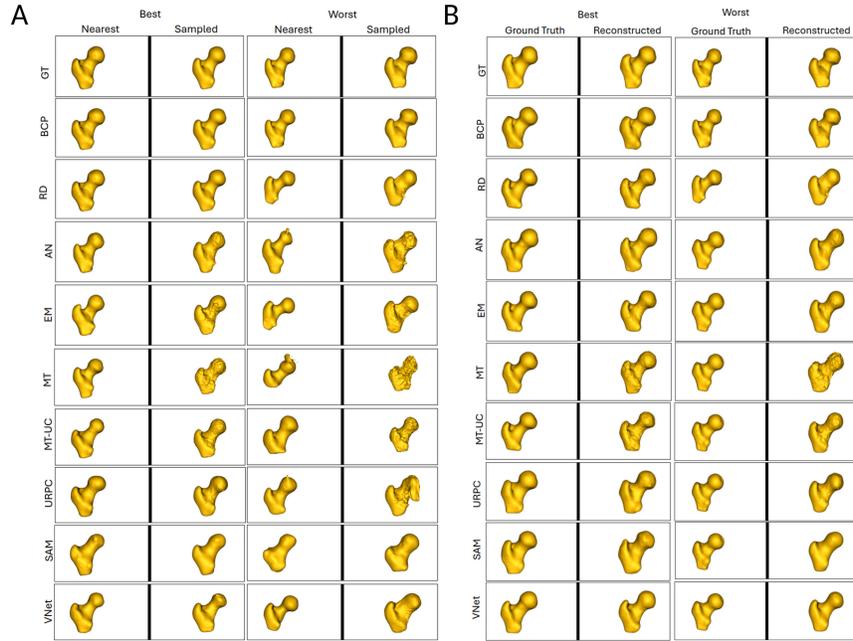
Fig. 5: **Specificity and Generalization examples for Femur (20%)** (A) Specificity - Reconstructions for the sampled shape and nearest ground truth shape to the sampled shape. Sampled shape from bidirectional copy-paste (BCP) are more closer to the nearest ground truth shape then other methods. (B) Generalization - Ground truth shape and shape reconstructions for unseen samples for best and worst case scenarios. After ground truth (GT), BCP has the better reconstructions
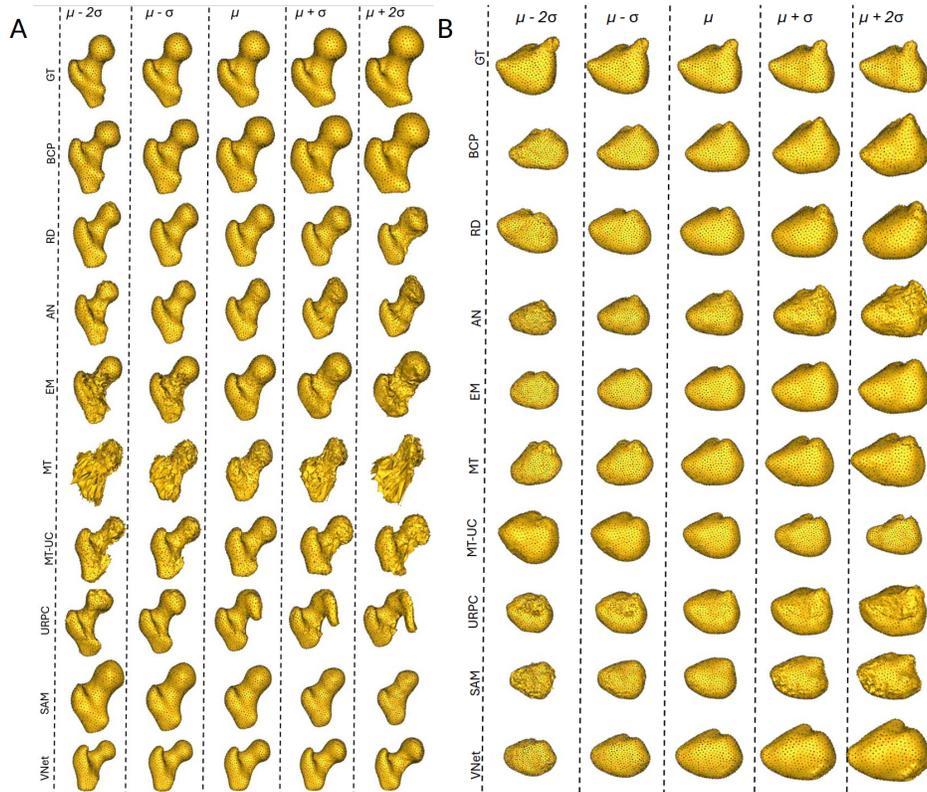
Fig. 6: **Second Mode Of variation Results for Femur and Left Atrium using 20% of labelled data.** We show first mode of variations for both Femur and Left Atrium datasets showing mean shape(μ), first ($\pm\sigma$) and second order ($\pm2\sigma$) variations from mean shape. We can clearly see that MT, MT-UC, URPC and EM fail to capture any relevant details for second mode of variations, where as V-Net capture as different mode of variation for both left atrium and femur. SAM and BCP capture same mode of variation for femur, but for left-atrium BCP and RD capture same mode of variation.