On the Viability of Semi-Supervised Segmentation Methods for Statistical Shape Modeling

Asma Khan*, Tushar Kataria*, Janmesh Ukey, and Shireen Elhabian

[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
(asmak,tushar.kataria,janmesh,shireen)@sci.utah.edu
* Equal Contribution

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 SSMs is often constrained by the necessity for expert-driven manual segmentation, a time-intensive and expensive process that restricts their broader utility. While deep learning approaches can estimate SSMs directly from unsegmented images, they still require segmentation annotations for SSM training. Although segmentations are unnecessary during deployment, the challenge of acquiring training manual annotations remains. Semi-supervised anatomical segmentation offers a promising solution to the annotation burden by leveraging a small labeled dataset alongside abundant unlabeled images. However, its effectiveness for downstream statistical shape model (SSM) construction remains largely unexplored. In this study, we systematically evaluate semi-supervised methods as alternatives to manual segmentation under low-annotation settings, using their predicted segmentations to construct SSMs and establish a comprehensive performance benchmark. Our findings reveal a clear divide in performance: while certain methods yield noisy segmentations that degrade SSM quality, others accurately capture the population's modes of variation comparable to those obtained from manual-segmentation SSMs despite a 60–80% reduction in manual annotation requirements. GitHub repository

Keywords: Statistical Shape Models · Semi-Supervised Segmentation · Modes of Variation · Annotation-Efficient SSM · Femur · Left Atrium

1 Introduction

Statistical shape models (SSMs) provide a quantitative framework for analyzing anatomical variations across populations, enabling the identification of normative trends and deviations and facilitating the development of diagnostic tools

and surgical planning systems [11]. 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, thereby bypassing the need for segmentation during inference [7,1,2,25,21,32,24,14,6,27,28,6,33,13,12,4]. However, these deep learning methods still necessitate anatomy segmentation to construct SSMs for training.

Automated deep learning-based anatomical segmentation can reduce the burden of generating segmentations for statistical shape model (SSM) construction. However, conventional supervised segmentation methods still require large amounts of manually annotated data for training. To address this limitation, numerous semi-supervised approaches have been proposed [11,20,5,35,26]. These methods leverage a small set of labeled volumes alongside a larger unlabeled dataset and employ techniques such as data augmentation [5], pseudo-labeling [5,26,35], consistency regularization [26,31], and entropy minimization [29] to achieve high-quality segmentations. Several methods (e.g., [18,34,23]) further exploit the inherent 3D structure of medical images, reducing annotation requirements from full-volume segmentations to a small number of annotated slices.

Despite the growing number of annotation-efficient segmentation methods, there are no established guidelines for assessing their suitability as alternatives to manual segmentation for statistical shape model (SSM) construction. Existing evaluations focus primarily on segmentation accuracy, which does not necessarily reflect the quality of the resulting SSMs. Predicted segmentations may achieve high overlap scores while altering anatomical boundaries and the statistical modes of shape variation. Therefore, evaluating these methods requires assessing the quality of the downstream SSMs rather than segmentation accuracy alone. To our knowledge, no comparable benchmark exists, making our evaluation a useful reference for developing medical SSMs with limited annotation (shown in Figure 1).

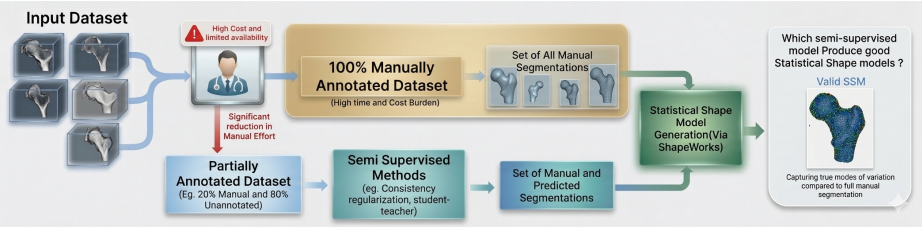

Fig. 1. Benchmarking pipeline. Semi-supervised segmentation models trained on limited labeled data and unlabeled images generate anatomical segmentations for SSM construction, which are benchmarked against SSMs from manual segmentations to assess the impact of reduced annotation on shape modeling.

In this paper, we present the first benchmark for evaluating whether semi-supervised segmentation methods can serve as annotation-efficient alternatives

to manual segmentation for statistical shape model (SSM) construction. Our benchmark compares SSMs generated from semi-supervised segmentations with those built from manual annotations, assessing their suitability for downstream population-level shape analysis under limited annotation settings. By identifying methods that preserve SSM quality while reducing annotation requirements, this work aims to improve the accessibility and practical adoption of SSMs in medical applications. The main contributions of this manuscript are:

- We establish the first benchmark for evaluating semi-supervised segmentation methods based on the quality of the statistical shape models they produce, using manual-segmentation-based SSMs as the reference.
- We evaluate SSM quality across varying annotation settings, including different proportions (20% and 40%) and annotation types (full-volume and sparse-slice) used to train semi-supervised segmentation models.
- We provide a comprehensive quantitative and qualitative analysis of downstream SSM performance, identifying the strengths and failure modes of current semi-supervised segmentation methods.

2 Methods

To determine whether semi-supervised methods can effectively replace manual annotations for SSM, we seek to address the following key questions:-

- *Do SSM models created using semi-supervised model predictions capture the same modes of variation compared to an SSM model using all manual segmentations?*
- *Are some semi-supervised methods more reliable than others?*
- *Does increasing the amount of annotated data for semi-supervised models improve SSM for the population cohort?*

Selected Semi-Supervised Methods. Semi-supervised medical image segmentation leverages partially labeled and all unlabeled data during training. We selected eight distinct methods that utilize labeled and unlabeled data in various ways, categorizing them into three distinct groups: (A) *Student-teacher methods*:- Mean Teacher (MT)[26] & Uncertainty-Aware Mean Teacher (UA-MT)[35], (B) *Pseudo-labeling methods:-* Mutual Correction Framework (MCF) [30], Orthogonal Annotation (DeSCO) [8], Bidirectional Copy-Paste (BCP)[5] and Correlation-Aware Mutual Learning (CAML) [10], and (C) *Multi-task learning methods:-* Dual-task Consistency (DTC) [8], and Shape-Aware Semi-supervised (SASSnet) [17]. A brief description of each of these methods is added below:

Mean Teacher (MT)[26]. Mean Teacher uses consistency regularization between *student-teacher* training paradigm. The *student* model is trained using the supervised loss on the labeled set, whereas the *teacher*'s parameters are updated as an exponential moving average of the student model. Consistency loss on unlabeled data aligns the prediction between the two models.

Uncertainty-Aware Mean Teacher (UA-MT)[35]. 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.

Mutual Correction Framework (MCF) [30]. MCF employs a student-teacher model paradigm to address cognitive biases learned during training, applying a rectification loss in regions where the model frequently makes errors. It then uses the dice scores of segmentations predicted by both models to determine the pseudo-labels for unlabeled data.

Bidirectional Copy-Paste (BCP)[5]. BCP integrates a bidirectional copy-paste framework into the Mean Teacher architecture. The student network receives input 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.

Correlation-Aware Mutual Learning (CAML) [10]. CAML introduces two modules: Cross-sample Mutual Attention, which uses transformers to compute attention between labeled and unlabeled samples within the same batch, and Omni-Correlation Consistency, a contrastive loss based on pixel-level latent space projections between labeled and unlabeled volumes. These modules enhance the transfer of information between labeled and unlabeled data, leading to improved segmentation performance.

Orthogonal Annotation (DeSCO) [8]. This paper presents a novel orthogonal annotation strategy for semi-supervised 3D medical image segmentation, where only three orthogonal slices are labeled and the remaining volumes are unlabeled. It employs cross-consistency and pseudo-labeling to enhance performance. This method significantly reduces the annotation effort required, making it more feasible for broader deployment.

Dual-task Consistency (DTC) [20]. Instead of applying consistency loss between labeled and unlabeled data, as seen in MT [26], UA-MT [35], and similar methods [31], DTC introduces a multi-task network to enforce consistency between two tasks. The chosen tasks are segmentation and level-set prediction, where the level-set function's zero level corresponds to the contour of the segmented anatomy. The model is trained to ensure consistent predictions across both tasks for both labeled and unlabeled data.

Shape-Aware Semi-supervised (SASSnet) [17]. SASSnet introduces a shape-aware semi-supervised strategy that leverages multitasking to predict both segmentation and the signed distance map (SDM) simultaneously. Additionally, it employs an adversarial loss between the SDM predictions for labeled and unlabeled data, which encourages the model to learn shape-aware features.

2.1 Experimental Design

To offer a thorough evaluation of semi-supervised methods for constructing Statistical Shape Models (SSM), we design two experimental strategies that enable detailed analysis and support the future deployment of these methods for end-user applications.

Consider a dataset $\mathcal{D}$, which consists of a collection of anatomical images ($\mathcal{I}$) along with their corresponding manual segmentations ($\mathcal{S}$). The dataset is split into training $\mathcal{D}_{Train}$ and testing set $\mathcal{D}_{test}$, such that $\mathcal{D}_{Train} \cap \mathcal{D}_{Test} = \emptyset$ and $\mathcal{D}_{Train} \cup \mathcal{D}_{Test} = \mathcal{D}$.

We define the SSM representation of an anatomy as Φ. The SSM is formulated as an optimization process aimed at minimizing the entropy of particles (or point clouds) on the surface of the anatomy [9]. This process is applied to all segmentations in the training dataset, denoted as $\mathcal{F}_{\Phi}(\cdot)$, where the function parameters consist of the segmentations and the initial particle locations.

Φ_{Train} establishes the best training SSM representation, as it utilizes all the manual annotations in the training set to construct an SSM, as described by the equation:

$$\Phi_{\text{train}} = \arg\min_{\Phi} \sum_{i \in D_{\text{train}}} F_{\Phi}(S_i; \phi_0), \tag{1}$$

where ϕ_0 represents the initial condition. ϕ_0 is set to origin for all particles. To generate the test SSM, we apply the same optimization over the test segmentations, with the Φ_{Train} as initial condition (fixed domain initialization [6]),

$$\Phi_{\text{test}} = \arg\min_{\Phi} \sum_{i \in D_{\text{test}}} F_{\Phi}(S_i; \Phi_{\text{train}}). \tag{2}$$

Φ_{Test} represents the manual-segmentation reference SSM on the test set, reflecting the optimal SSM representation achievable using all manual annotations. However, since these test samples are unseen by the SSM model, distribution shifts between the training and test segmentations—such as variations in size and contour—can result in generalization errors.

Let us denote all selected semi-supervised methods discussed above as a set called $SEMI \in \{MT, UA\text{-}MT, BCP, CAML, DeSCO, DTC, MCF, SASSnet\}$. The predicted segmentations from these methods can be represented by $\tilde{\mathcal{S}}^k \forall k \in SEMI$. To evaluate the potential of semi-supervised methods as an alternative to manual segmentations in SSM construction, we investigate two strategies:

- Φ_{Train} **available (Strategy 1)**: Instead of using manual segmentations as in equation 2, predicted segmentations from each semi-supervised method ($k \in SEMI$) on the test set are used to obtain the SSM representation as,

$$\widetilde{\Phi}^k_{\text{test}} = \arg\min_{\Phi} \sum_{i \in D_{\text{test}}} F_{\Phi}(\widetilde{S}^k_i; \Phi_{\text{train}}), \qquad k \in \text{SEMI}. \tag{3}$$

  compared to Φ_{Test}, $\tilde{\Phi}^k_{Test}$ includes errors from both distribution shifts between the training and test segmentations and the absence of manual annotations for test set.
- $\mathbf{\Phi}_{Train}$ **not available (Strategy 2)**: This strategy further limits access to manual segmentations by assuming that only limited training annotations are available for SSM representation on the training set. Only a small amount of segmentation data (20% or 40% of manual segmentations) are provided to

train the semi-supervised methods, which then generate predicted segmentations for both the training and test sets. For each semi-supervised method ($k \in SEMI$) we obtain the SSM representation of the training set as:

$$\widetilde{\Phi}^k_{\text{train}} = \arg\min_{\Phi} \sum_{i \in D_{\text{train}}} F_{\Phi}(\widetilde{S}^k_i; \phi_0), \qquad k \in \text{SEMI}. \tag{4}$$

where ϕ_0 is the initial condition. Using $\tilde{\Phi}^k_{Train}$ representation, we create test SSM as,

$$\widetilde{\Phi}^k_{\text{test}} = \arg\min_{\Phi} \sum_{i \in D_{\text{test}}} F_{\Phi}(\widetilde{S}^k_i; \widetilde{\Phi}^k_{\text{train}}), \qquad k \in \text{SEMI}. \tag{5}$$

compared to Φ_{Test}, the $\tilde{\Phi}^k_{Test}$ representation is impacted by the same errors from Strategy 1, with the added impact due to absence of manual segmentations on the training set. The two strategies help isolate different sources of error: Strategy 1 isolates the generalization error only, whereas Strategy 2 captures both the generalization error and the impact of the unavailability of manual segmentations for constructing the training SSM and testing SSM.

2.2 Datasets and Metrics for Performance Evaluation

Datasets. We evaluated two datasets: 59 volumes from the public CARMA NAMIC left-atrium (LA) dataset [22], acquired using LGE-MRI, and 49 volumes from an in-house FEMUR dataset [6] acquired using CT. The FEMUR dataset was resampled to isotropic 0.5 x 0.5 x 0.5 mm voxel spacing while NAMIC reports the resolution as 0.625 x 0.625 x 2.5 mm. To assess semi-supervised performance, we tested two levels of labeled data availability (20% and 40%) during training. The FEMUR dataset (49 volumes) was split into 40 training samples (20%: 8 labeled/32 unlabeled; 40%: 16 labeled/24 unlabeled) and 9 testing samples. Similarly, the NAMIC dataset was split into 50 training samples (20%: 10 labeled/40 unlabeled; 40%: 20 labeled/30 unlabeled) and 9 testing samples. We selected the 20% setting to evaluate a highly annotation-constrained regime using no more than 10 manually labeled training volumes, and the 40% setting to test whether additional labeled data improves both semi-supervised segmentation performance and downstream SSM quality.

Implementation Details. We used the original codebases with default parameters for all semi-supervised methods. Models were trained on a 12GB NVIDIA 1080 Ti GPU, except for CAML, which required a 40GB NVIDIA A100 due to memory constraints. ShapeWorks [9] was utilized to generate all shape models for analysis given its established efficacy [11]. All SSMs were constructed using ShapeWorks v6.5.1 with 1024 correspondence particles per shape. Grooming used an isovalue of 0.5 and 30 antialiasing iterations. Particle optimization used 1000 iterations per split and 500 optimization iterations, with regularization decreasing from 1000 to 10 and recomputed every two iterations;

Procrustes alignment was enabled. All primary experiments used the same predefined train/test split and a single training run per configuration with a fixed random seed of 1337.

Evaluation Metrics. We evaluate segmentation accuracy using Dice and Jaccard scores, Average Surface Distance (ASD), and the 95th-percentile Hausdorff distance [26,35,15,30,16]. We report HD95 instead of the maximum Hausdorff distance to reduce sensitivity to isolated outlier voxels while preserving sensitivity to boundary errors.

In an ideal statistical shape model (SSM), point locations and neighbourhood relationships are preserved across the segmentations of the anatomy of interest for all patients. This enables meaningful analysis of shape variation across the population. In Statistical Shape Modeling (SSM), aligned corresponding points from all shapes are averaged to create a mean shape representing the typical anatomy. Principal Component Analysis (PCA) is then applied to identify the main modes of shape variation, with each mode describing an independent pattern of anatomical change. These modes can be used to compare healthy and diseased populations, test medical hypotheses, and visualize shape differences by deforming the mean shape along each principal component. Therefore to compare SSM from manual segmentation and predicted segmentation from semi-supervised methods, we compute four metrics based on the PCA projections of SSM particles [9,4,6,3]:

- **Compactness:** Measured using the area under the curve (AUC). Higher AUC values indicate a more efficient SSM, requiring fewer principal component analysis (PCA) modes (parameters) to represent a given amount of shape variation in the training set.
- **Specificity:** The specificity metric measures the extent to which an SSM generates anatomically plausible instances that belong to the shape class represented by the training set. It is computed as the average distance between shapes randomly generated by the SSM and their closest corresponding shapes in the training set.
- **Generalization:** Assesses transferability to unseen subjects using the distance between test particle correspondences and their PCA reconstructions (lower is better).
- **Grassmannian distance [19]:** We introduce this as a novel evaluation method in this application to measure the distance between two PCA subspaces for each mode (lower is better).

Statistical Shape Models (SSMs) are also assessed qualitatively by comparing their first two modes of variation.

3 Results and Discussion

Segmentation Performance of Semi-Supervised methods. We report the segmentation performance on test data in Tables 1 and 2, evaluating each semi-supervised method using 20% and 40% labeled training splits. The reported

standard deviations quantify inter-subject variability on the held-out test set. Several noteworthy observations emerge from these results:

(1) *Scaling with Annotated Data*: Increasing the amount of annotated training data does not universally guarantee improved performance. While methods like MCF scale exceptionally well, DeSCO and MT exhibit a slight degradation in segmentation accuracy when transitioning from 20% to 40% labeled data. DeSCO, in particular, shows severe boundary instability on the LA dataset at 40% data, where its HD95 more than doubles.

(2) *Methodological Superiority*: MCF emerges as the strongest overall model, particularly at the 40% data split—closely followed by BCP, which excels at the 20% split. Conversely, purely Student-Teacher architectures like MT and UA-MT consistently rank among the lowest-performing methods. Overall, architectures utilizing pseudo-labeling (MCF, BCP, CAML) tend to outperform other paradigms on these tasks.

Table 1. Left Atrium Semi-Supervised Segmentation Results. Metrics include Dice, Jaccard (Jacc.), ASD, and HD95 on test data, reported as $\text{mean}_{\text{standard deviation}}$ across nine held-out cases. ↑/↓ indicate higher/lower is better, and the best method per configuration is bolded.

Method	20% Data (10 labeled)				40% Data (20 labeled)			
	Dice (↑)	**Jacc.** (↑)	**ASD** (↓)	**HD95** (↓)	**Dice** (↑)	**Jacc.** (↑)	**ASD** (↓)	**HD95** (↓)
DeSCO	$0.862_{0.022}$	$0.757_{0.033}$	$1.81_{0.54}$	$7.17_{2.18}$	$0.841_{0.047}$	$0.728_{0.069}$	$4.51_{1.74}$	$15.54_{6.31}$
SASS	$0.855_{0.029}$	$0.748_{0.045}$	$2.16_{0.83}$	$8.92_{3.12}$	$0.886_{0.022}$	$0.795_{0.035}$	$1.42_{0.37}$	$5.61_{1.18}$
DTC	$0.861_{0.030}$	$0.757_{0.046}$	$1.83_{0.77}$	$7.80_{2.74}$	$0.881_{0.026}$	$0.789_{0.042}$	$1.49_{0.52}$	$6.19_{2.71}$
MCF	$0.868_{0.028}$	$0.768_{0.045}$	$2.01_{0.85}$	$7.78_{3.82}$	$\mathbf{0.902}_{0.016}$	$\mathbf{0.823}_{0.027}$	$1.34_{0.33}$	$\mathbf{4.62}_{1.70}$
CAML	$0.889_{0.016}$	$0.800_{0.025}$	$\mathbf{1.32}_{0.27}$	$6.18_{1.76}$	$0.893_{0.018}$	$0.807_{0.030}$	$\mathbf{1.30}_{0.35}$	$6.12_{1.56}$
MT	$0.869_{0.024}$	$0.770_{0.038}$	$2.02_{0.81}$	$7.85_{2.78}$	$0.859_{0.035}$	$0.755_{0.054}$	$1.59_{0.40}$	$7.30_{2.72}$
UA-MT	$0.872_{0.025}$	$0.774_{0.038}$	$2.94_{2.88}$	$10.40_{10.39}$	$0.873_{0.030}$	$0.776_{0.046}$	$1.68_{0.83}$	$6.66_{3.47}$
BCP	$\mathbf{0.896}_{0.025}$	$\mathbf{0.813}_{0.040}$	$1.53_{0.31}$	$\mathbf{5.87}_{1.67}$	$0.901_{0.016}$	$0.821_{0.026}$	$1.34_{0.22}$	$4.71_{1.37}$

(3) *Volume vs. Boundary Accuracy*: The results underscore the necessity of boundary metrics. While overlap scores (Dice) can appear tight, surface distances (ASD and HD95) expose severe localization errors. For example, on the 20% Femur dataset, MT achieves a respectable 0.931 Dice score but exhibits substantially larger HD95 (15.52) compared to BCP (2.05). CAML and BCP consistently yield low boundary errors, while MCF demonstrates the most significant boundary refinement when trained with 40% labeled data.

These segmentation results establish which methods produce accurate masks, but they do not determine whether the resulting surfaces preserve the population-level shape space needed for SSM analysis.

3.1 Statistical Shape Modeling Performance Using Different Semi-Supervised Methods

Quantitative results for Strategy 1. Figure 2 presents the quantitative evaluation of Strategy 1, where the statistical shape model (SSM) is trained using

Table 2. Femur Semi-Supervised Segmentation Results. Metrics include Dice, Jaccard (Jacc.), ASD, and HD95 on test data, reported as $\text{mean}_{\text{standard deviation}}$ across nine held-out cases. ↑/↓ indicate higher/lower is better, and the best method per configuration is bolded.

Method	20% Data (8 labeled)				40% Data (16 labeled)			
	Dice (↑)	**Jacc.** (↑)	**ASD** (↓)	**HD95** (↓)	**Dice** (↑)	**Jacc.** (↑)	**ASD** (↓)	**HD95** (↓)
DeSCO	$0.954_{0.009}$	$0.913_{0.016}$	$0.92_{0.24}$	$2.60_{0.71}$	$0.947_{0.014}$	$0.899_{0.025}$	$1.40_{0.81}$	$5.19_{4.24}$
SASS	$0.950_{0.013}$	$0.905_{0.024}$	$1.58_{0.83}$	$5.14_{4.51}$	$0.958_{0.009}$	$0.920_{0.016}$	$0.77_{0.22}$	$2.48_{0.69}$
DTC	$\mathbf{0.961}_{0.012}$	$\mathbf{0.925}_{0.022}$	$0.76_{0.33}$	$2.38_{1.16}$	$0.961_{0.009}$	$0.924_{0.017}$	$0.67_{0.17}$	$2.31_{0.94}$
MCF	$0.959_{0.025}$	$0.922_{0.044}$	$1.57_{1.28}$	$3.99_{4.43}$	$\mathbf{0.974}_{0.007}$	$\mathbf{0.950}_{0.012}$	$\mathbf{0.53}_{0.21}$	$\mathbf{1.49}_{0.51}$
CAML	$0.960_{0.010}$	$0.923_{0.012}$	$0.84_{0.22}$	$2.86_{1.15}$	$0.964_{0.006}$	$0.931_{0.011}$	$0.69_{0.12}$	$2.46_{0.58}$
MT	$0.931_{0.019}$	$0.871_{0.034}$	$3.99_{2.22}$	$15.52_{11.20}$	$0.923_{0.060}$	$0.863_{0.095}$	$2.32_{2.64}$	$6.70_{8.45}$
UA-MT	$0.944_{0.017}$	$0.894_{0.030}$	$2.73_{1.38}$	$9.67_{8.33}$	$0.946_{0.019}$	$0.898_{0.034}$	$1.93_{1.93}$	$7.29_{10.28}$
BCP	$\mathbf{0.962}_{0.004}$	$\mathbf{0.926}_{0.007}$	$\mathbf{0.68}_{0.05}$	$\mathbf{2.05}_{0.09}$	$0.964_{0.016}$	$0.930_{0.026}$	$0.66_{0.22}$	$2.00_{1.37}$

manually segmented data, while semi-supervised segmentations are used to construct the test SSM. The resulting test SSM is then compared with the corresponding test SSM generated from manual segmentations. Since both approaches use the same training SSM, the evaluation focuses only on the Generalization and Grassmannian distance metrics.

The generalization error measures how closely the SSM constructed from semi-supervised segmentations approximates the ground-truth SSM obtained from manual segmentations. Therefore, a lower generalization error indicates that the semi-supervised segmentations produce shape representations that are more consistent with the ground truth.

The Grassmannian distance evaluates the similarity between the PCA subspaces of the manually generated and semi-supervised test SSMs. As a dimensionless metric, smaller values indicate greater similarity between the learned shape spaces. However, while a low Grassmannian distance suggests that the underlying PCA subspaces are closely aligned, it does not necessarily imply that the two models will exhibit equivalent shape reconstruction or generalization performance.

For the FEMUR dataset, test SSMs constructed from manual segmentations (GT) consistently achieve lower generalization errors across all PCA modes than those constructed from semi-supervised predictions, irrespective of whether 20% or 40% of the training data is annotated. Among the semi-supervised methods trained with 20% annotations, BCP most closely matches the GT model in terms of generalization performance, followed by CAML. The Grassmannian distance analysis shows that BCP also produces the closest PCA subspace to the GT-based SSM, with CAML ranking second. When 40% of the training data is annotated, MCF improves its generalization performance and becomes the second-best method after GT, while BCP ranks third. Nevertheless, BCP continues to exhibit the lowest Grassmannian distance, with CAML and MCF closely following. Furthermore, the overall Grassmannian distances decrease when using 40% annotated data, indicating that increasing annotation availability improves the agreement between the PCA subspaces learned from semi-supervised and manually segmented test SSMs.

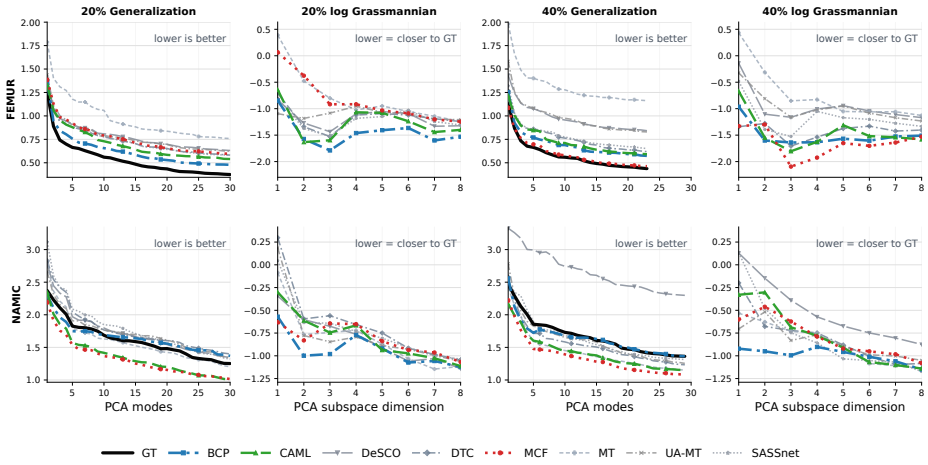

Fig. 2. Strategy 1: downstream SSM quality when the manual training SSM is available. Generalization error across PCA modes and log Grassmannian distance to the manual-segmentation SSM subspace are shown for SSMs built from semi-supervised test segmentations and initialized from the manual-segmentation training SSM. Lower values are indicative of better agreement. The generalization curves are indexed by reconstruction PCA mode, whereas the Grassmannian curves compare (k)-dimensional PCA subspaces constructed from the test shapes. Because each test SSM contains nine subjects, its PCA rank is at most eight.

The NAMIC dataset exhibits a different trend. Several semi-supervised methods achieve lower generalization errors than the GT-based SSM, particularly when trained with 20% annotated data, and this behavior persists for most methods when 40% of the data is annotated. However, these improvements in generalization do not imply that the learned statistical shape models recover the same population-level shape statistics as the manual-segmentation reference. The Grassmannian distance remains comparable to that observed for the FEMUR dataset, indicating that although the semi-supervised SSMs reconstruct test shapes effectively, their PCA subspaces remain distinct from the GT SSM.

Generalization measures how accurately an SSM reconstructs unseen test correspondences, whereas the Grassmannian distance quantifies the similarity between the PCA subspaces of two SSMs. Consequently, a lower generalization error does not necessarily indicate recovery of the same anatomical modes of variation. Instead, the improved generalization observed for some semi-supervised methods may result from smoother or more internally consistent predicted segmentations, reduced annotation variability, or differences in the distributions of the training and test surfaces. Such behavior is expected for many semi-supervised approaches that incorporate entropy minimization or consistency regularization, which encourage consistent predictions across the dataset while potentially altering the underlying statistical shape basis.

To evaluate whether segmentation accuracy predicts downstream SSM quality, Fig. 3 compares Dice and ASD with Strategy 1 Grassmannian distance. Strong correlations can be observed for the femur but not for the left atrium, indicating that segmentation accuracy is not a consistent predictor of downstream shape-space quality across anatomies and that the relationship is anatomy-dependent.

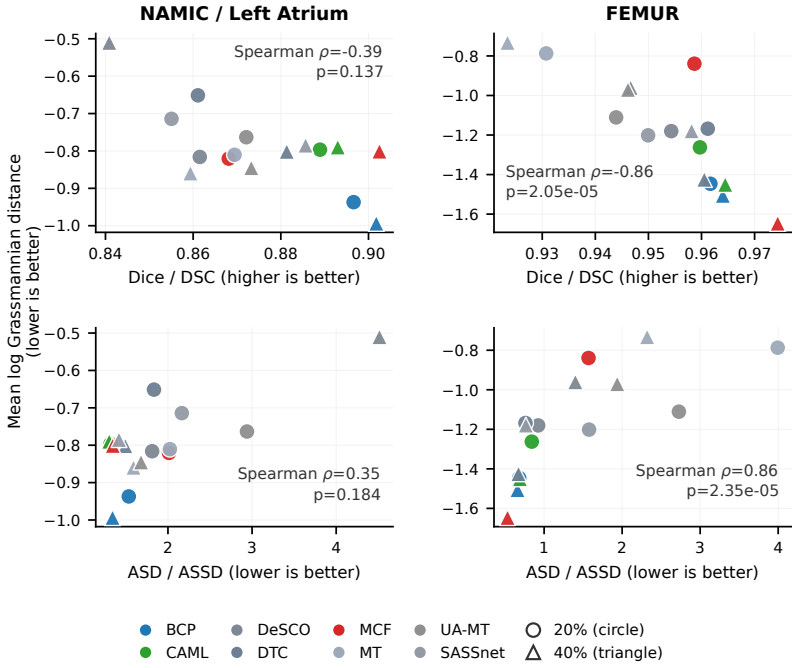

Fig. 3. Segmentation accuracy versus downstream SSM subspace agreement. Each point represents a semi-supervised method at one annotation setting, with circles for 20% and triangles for 40% labeled data. Dice/DSC and ASD/ASSD are plotted against the mean log Grassmannian distance from Strategy 1; higher Dice is better, while lower ASD and Grassmannian distance indicate better boundary accuracy and closer agreement with the manual SSM subspace.

Quantitative results for Strategy 2. Figure 4 evaluates Strategy 2, the stricter deployment setting in which both training and test SSMs are constructed from semi-supervised predictions. It reports compactness, specificity, generalization and Grassmannian distance for FEMUR and NAMIC datasets for models trained with 20% and 40% of annotations.

For FEMUR, the ground truth SSM achieves the highest AUC for compactness, indicating that it is the most compact model, followed closely by CAML and BCP for both 20% and 40% annotated models. Generalization scores show a similar trend, with the GT SSM consistently performing best, followed by BCP and DTC as the second and third best models for 20% annotation, while MCF

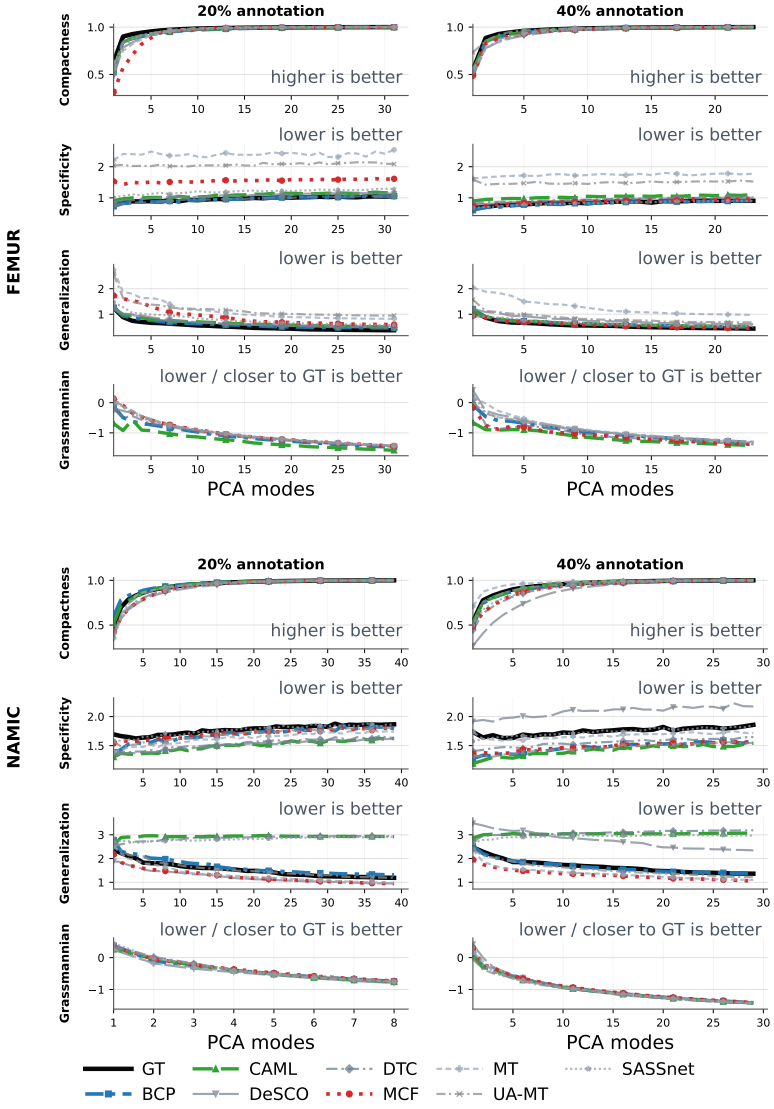

Fig. 4. Strategy 2: downstream SSM quality when both training and test SSMs are built from semi-supervised segmentations. Compactness, specificity, generalization, and log Grassmannian distance are shown for FEMUR and NAMIC under 20% and 40% labeled-data settings. Higher compactness indicates a more efficient shape model; lower specificity, generalization, and Grassmannian distance indicate better shape fidelity or closer agreement with the manual-segmentation reference.

and DTC rank second and third for 40% annotation. In terms of specificity, DTC closely matches the GT SSM, with DeSCO ranking third for 20% models. For 40% models, SASSnet and BCP produce specificity scores that are also very

similar to the GT SSM. The Grassmannian distance remains comparable to that observed in FEMUR results from Strategy 1, indicating that the PCA subspace of test SSMs—whether initialized with the GT train SSM or the predicted train SSM—remains similarly distant.

For NAMIC, results align with Strategy 1, where the GT model does not achieve the best generalization performance. Instead, DeSCO, MCF, and DTC emerge as the top-performing models in generalization. The GT model is also not the most compact; BCP performs best for 20% annotated models, while DTC ranks highest for 40% models. A similar pattern is observed for specificity scores. Unlike FEMUR, where increasing annotated data has little impact—as reflected by the unchanged Grassmannian distance—NAMIC shows a slight improvement in Grassmannian distance when moving from 20% to 40% annotation, suggesting that additional annotations provide a small but positive effect on SSM subspace alignment.

3.2 Qualitative Analysis

Figures 5 and 6 provide representative qualitative examples of the first two modes of variation under the most annotation-limited setting. We show the 20% setting because it is the most annotation-constrained method in our setting and therefore most clearly exposes shape-model failure modes.

Qualitative results for Strategy 1. Figure 5 illustrates representative first and second modes of variation for Strategy 1. The results demonstrate that nearly all methods generate smooth and interpretable statistical shape models (SSMs), effectively capturing both the first and second modes. However, in some cases, the direction of variation is reversed in the test SSMs produced by semi-supervised models. Specifically, BCP, MT, and UA-MT exhibit this reversal for both NAMIC and FEMUR in modes 1 and 2, while SASSnet also shows a reversed mode of variation for NAMIC in mode 1. Despite these reversals, most models generate qualitatively smooth, meaningful, and interpretable representations of dominant shape variation.

Qualitative results for Strategy 2. Figure 6 illustrates representative first and second modes of variation for Strategy 2. From these figures, we observe that SSMs generated for the FEMUR dataset by BCP, CAML, DeSCO, and DTC exhibit smooth shapes, closely resembling the ground truth SSM, while other methods result in poor-quality SSMs. Additionally, CAML, DeSCO, and DTC effectively capture the first mode of variation, though are noticeable differences in BCP's results. A similar trend is observed for NAMIC, where the same semi-supervised methods also produce smooth SSMs. However, none of the SSM models successfully capture the pulmonary vein ostia (left atrial protrusions), as the produced SSMs exhibit overly smooth and poorly defined anatomical structures, leading to significantly noticeable discrepancies in both mode 1 and mode 2 compared to the GT SSM.

Practical guidelines for usage of semi-supervised methods for SSMs. Segmentation metrics alone were not sufficient to predict downstream SSM quality. Although lower ASD and HD95 generally correlated with improved shape

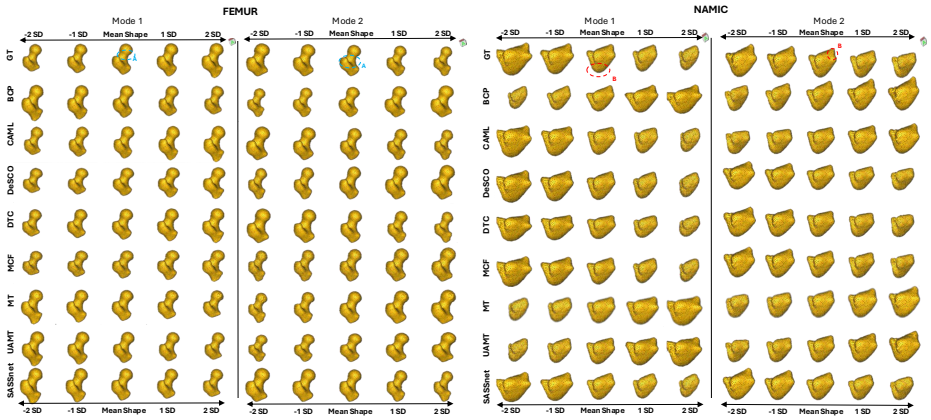

Fig. 5. Strategy 1 qualitative mode analysis. Representative first and second modes of variation for FEMUR and left atrium SSMs, using the manual training SSM and test SSMs from semi-supervised segmentations with 20% labeled data. Each row shows shapes at -2σ, -1σ, mean, $+1\sigma$, and $+2\sigma$. Annotation A highlights femoral variation near the head–neck region, while annotation B highlights thin left-atrial structures near the pulmonary vein/ostial region. Highlighted regions are illustrative and do not imply other regions are unchanged.

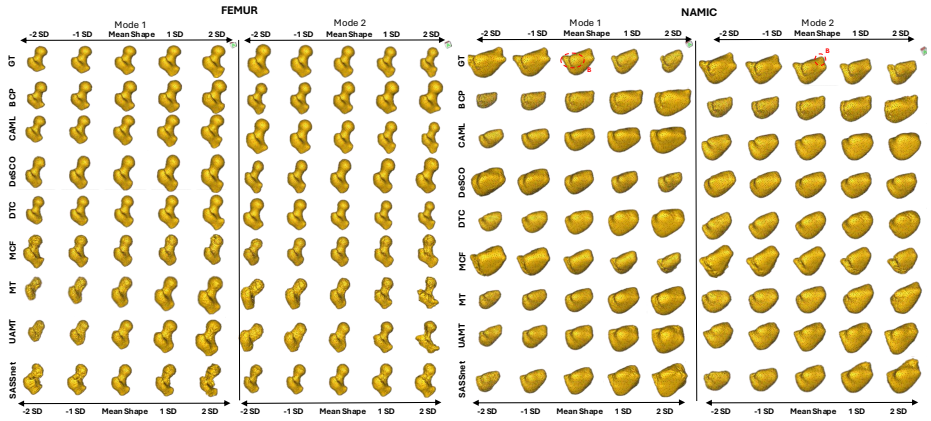

Fig. 6. Strategy 2 qualitative mode analysis. Representative first and second modes of variation for FEMUR and left atrium SSMs, with both training and test SSMs constructed from semi-supervised segmentations using 20% labeled data. Each row shows shapes at -2σ, -1σ, mean, $+1\sigma$, and $+2\sigma$. Annotation B highlights thin left-atrial structures near the pulmonary vein/ostial region, where differences in smoothness and anatomical detail are most visible. Highlighted regions are illustrative and do not imply other regions are unchanged.

modeling, methods with the lowest boundary errors did not always produce the most accurate SSMs. Likewise, models with competitive Dice scores or lower

generalization error did not necessarily recover the PCA subspace derived from manual segmentations. Methods incorporating boundary refinement and pseudo-label correction, such as BCP, MCF, and CAML, more consistently preserved shape-space quality, but no single method performed best across all anatomies and annotation budgets.

Qualitative analysis further showed that, despite achieving superior segmentation accuracy, methods such as CAML, BCP, and MCF did not always reproduce the same principal modes of variation as models built from manual segmentations. This suggests that optimizing for segmentation consistency alone may not be sufficient to preserve the true anatomical variability of a population. A promising direction for future semi-supervised methods is to explicitly model population-level shape variation, for example through statistical shape priors, distribution-aware learning, or conditioning segmentation on anatomical and image-texture variability. Such approaches may better preserve the underlying distribution of organ shapes while maintaining segmentation accuracy.

Overall, these findings highlight the importance of validating semi-supervised segmentation using downstream shape-space metrics before applying it to clinical morphology studies or shape-based biomarkers. They also suggest that semi-supervised segmentations may be more suitable for constructing test SSMs than training SSMs. While test models primarily evaluate new shapes against an existing statistical model, training SSMs must accurately capture the full distribution of anatomical variability. The observed loss of agreement in the principal modes of variation indicates that current semi-supervised methods may not faithfully preserve this true variability, potentially limiting their suitability for building population-representative training SSMs.

Limitations. This study evaluated two anatomies using default implementations of representative semi-supervised methods. Results may differ with architecture tuning, alternative annotation protocols, or anatomies with more complex topology. Because accurate shape-space reconstruction is a prerequisite for morphology-based analysis, this work focused on SSM quality rather than clinical prediction. The datasets are modest in size, with 49 FEMUR volumes and 59 NAMIC volumes, which limits the statistical power of method comparisons and motivates validation on larger sample sizes as part of future work which will also investigate whether differences in PCA subspace alignment translate to improved disease classification, longitudinal progression modeling, and other clinical applications. Additionally, without a fully supervised baseline, we treat the 40% annotation condition as a proxy for determining the benefit of approaching unlabeled data.

4 Conclusion

We presented a benchmark for evaluating whether semi-supervised segmentation methods can replace manual segmentations for statistical shape modeling. Our results show that annotation-efficient segmentation can substantially reduce manual labeling requirements, but downstream SSM quality depends strongly

on the semi-supervised method, anatomy, and evaluation metric. In particular, generalization and segmentation overlap metrics can obscure changes in the learned shape subspace, motivating the use of Grassmannian subspace analysis and qualitative comparison of PCA modes of variation as complementary measures of SSM quality. These findings provide practical guidance for deploying semi-supervised segmentation in shape-driven population studies and suggest that current semi-supervised methods, while effective at improving segmentation consistency, do not necessarily preserve the underlying population-level anatomical variability. Future semi-supervised approaches should therefore incorporate mechanisms that explicitly preserve shape distributions and population statistics, enabling segmentations that more faithfully reflect true anatomical variation while maintaining high segmentation accuracy.

Data and Code Availability. The NAMIC CARMA dataset is publicly available through the NA-MIC Downloads repository [22]. The in-house FEMUR data cannot be publicly released and can be requested by contacting ShapeWorks [9] support at the University of Utah. GitHub repository

Ethics Statement. The FEMUR CT dataset was collected under the approval of the Institutional Review Board (IRB), approval number 00056086.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Acknowledgments.The National Institutes of Health supported this work under grant numbers NIBIB-U24EB029011 and NIAMS-R01AR076120. The content is solely the authors' responsibility and does not necessarily represent the official views of the National Institutes of Health.

References

1. 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)
2. Adams, J., Elhabian, S.: From images to probabilistic anatomical shapes: A deep variational bottleneck approach. arXiv preprint arXiv:2205.06862 (2022)
3. Adams, J., Elhabian, S.: Point2ssm: learning morphological variations of anatomies from point clouds. In: International Conference on Learning Representations. vol. 2024, pp. 13493–13513 (2024)
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. Cai, H., Li, S., Qi, L., Yu, Q., Shi, Y., Gao, Y.: Orthogonal annotation benefits barely-supervised medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3302–3311 (2023)
9. 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)
10. Gao, S., Zhang, Z., Ma, J., Li, Z., Zhang, S.: Correlation-aware mutual learning for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 98–108. Springer (2023)
11. 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)
12. Iyer, K., Adams, J., Elhabian, S.Y.: Scorp: Statistics-informed dense correspondence prediction directly from unsegmented medical images. arXiv preprint arXiv:2404.17967 (2024)
13. 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)
14. 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)
15. Kataria, T., Knudsen, B., Elhabian, S.: To pretrain or not to pretrain? a case study of domain-specific pretraining for semantic segmentation in histopathology. In: Workshop on Medical Image Learning with Limited and Noisy Data. pp. 246–256. Springer (2023)
16. Kataria, T., Knudsen, B.S., Elhabian, S.Y.: Unsupervised domain adaptation for semantic segmentation under target data scarcity. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2024)
17. Li, S., Zhang, C., He, X.: Shape-aware semi-supervised 3d semantic segmentation for medical images. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23. pp. 552–561. Springer (2020)
18. 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)
19. 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
20. Luo, X., Chen, J., Song, T., Wang, G.: Semi-supervised medical image segmentation through dual-task consistency. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 8801–8809 (2021)
21. 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)
22. National Alliance for Medical Image Computing: About NA-MIC. https://na-mic.org (2026), accessed: 2026-08-18

23. Nihalaani, R., Kataria, T., Adams, J., Elhabian, S.Y.: Estimation and analysis of slice propagation uncertainty in 3d anatomy segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 273–285. Springer (2024)
24. 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)
25. 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)
26. 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)
27. Ukey, J., Elhabian, S.: Localization-aware deep learning framework for statistical shape modeling directly from images. In: Medical Imaging with Deep Learning (2023)
28. 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
29. 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)
30. Wang, Y., Xiao, B., Bi, X., Li, W., Gao, X.: Mcf: Mutual correction framework for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15651–15660 (2023)
31. 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)
32. 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)
33. Xu, H., Elhabian, S.Y.: Image2ssm: Reimagining statistical shape models from images with radial basis functions. arXiv preprint arXiv:2305.11946 (2023)
34. Yeung, P.H., Namburete, A.I., Xie, W.: Sli2vol: Annotate a 3d volume from a single slice with self-supervised learning. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24. pp. 69–79. Springer (2021)
35. 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 Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22. pp. 605–613. Springer (2019)