Learning a Sampling-Free Variational DNN Plugin from Tiny Training Sets to Refine OOD Segmentation With Uncertainty Estimation

Jimut B. **Pal** [1], Suyash P. **Awate** [1,2]

1 Centre for Machine Intelligence and Data Science (C-MInDS), Indian Institute of Technology (IIT) Bombay, Mumbai
2 Computer Science and Engineering (CSE) Department, Indian Institute of Technology (IIT) Bombay, Mumbai

Abstract
Deep neural networks (DNNs) frequently fail to generalize to out-of-distribution (OOD) medical images because of variations in scanners and acquisition protocols. Retraining DNN models to address these distribution shifts is often impractical due to the high cost of acquiring and annotating new medical datasets. To address this, we introduce VarDeepPCA, a novel lightweight variational DNN framework designed to restore/refine degraded segmentation maps by leveraging intrinsic geometric priors. Unlike existing approaches that require target-domain data or extensive pre-training, our VarDeepPCA explicitly learns a distribution of valid anatomical geometries using only small in-distribution (ID) datasets. Theoretically, our novel variational learning framework leverages a reinterpretation of the softmax mapping to implicitly perform exact distribution modeling, thereby enabling computationally efficient, sampling-free learning and inference. This also enables VarDeepPCA to provide uncertainty estimates associated with its restored segmentation maps. We empirically validate our framework across 4 distinct clinical applications, using 14 publicly available datasets, involving segmentation of the myocardium, neuroretinal rim, prostate, and fetal head. Comparisons against 15 existing methods demonstrate that VarDeepPCA consistently restores segmentation maps produced by the existing methods on OOD data to (i) significantly improve anatomical plausibility of geometries and clinical utility of the segmentations, and (ii) significantly reduce errors, without needing any more training data than that used by existing methods.

Keywords
Out-of-distribution images, segmentation refinement, plugin, geometric prior learning, small training set, sampling-free variational learning, uncertainty.

Article informations

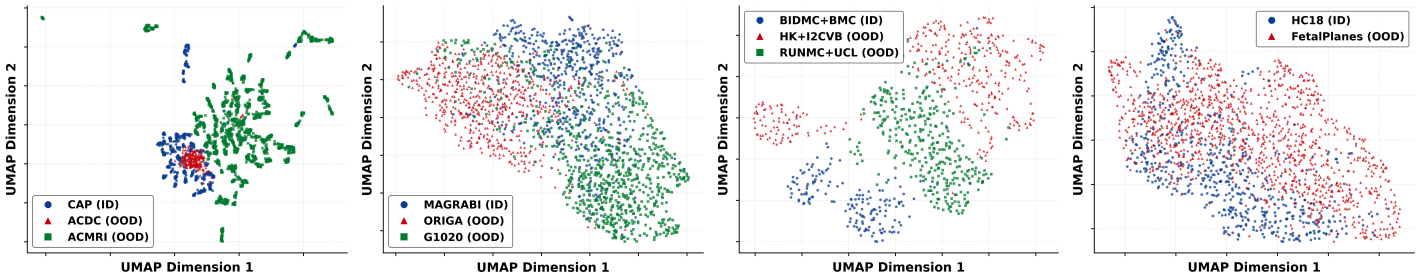
Corresponding author: pal.jimut@iitb.ac.in

1. Introduction

Deep neural networks (DNNs) exhibit significant performance degradation when applied to out-of-distribution (OOD) data (Farquhar and Gal, 2022; Tran et al., 2020). Our study addresses OOD data as it appears in real-world clinical scenarios. We define *OOD images* to be of the same anatomical object present in our training/in-distribution (ID) images but acquired from hospitals or scanners different from those associated with the training images. In this cross-site setting, the inevitable variations in imaging protocols, devices, and reconstruction schemes give rise to *distribution shifts* (Liang et al., 2025; Karani et al., 2021; Pal and Awate, 2024b), as seen in Figure 1, posing a critical barrier to the *clinical deployment* of DNNs. OOD data are assumed (Farquhar and Gal, 2022; Tran et al., 2020; Liang et al., 2025; Karani et al., 2021; Pal and Mj, 2023) to be unavailable during DNN training; retraining is often impractical due to the high cost of acquiring and annotating new data.

We consider segmentation tasks with a single object of interest per image, i.e., segmenting foreground versus background; our framework may also be extended to multi-class segmentation problems involving multiple kinds of objects of interest in an image. We propose a method to correct the degraded outputs of existing DNNs on OOD data by leveraging *priors* on the *inherent geometry* of the object of interest, which remains largely invariant to the aforementioned OOD variations. Let a *segmentation map* be an image where each pixel value lies within $[0, 1]$, representing the probability of that pixel belonging to the object of interest. Our novel DNN framework learns the principal modes of variation from ID segmentation maps to explicitly model a distribution on the manifold of valid anatomical object geometries. This learned low-dimensional distribution then serves as a powerful geometrical prior to rectify erroneous segmentations produced by existing DNNs on OOD images.

We apply our method to four diverse medical appli-



(a) Cardiac MRI (b) Retinal Images (c) Prostate MRI (d) Fetal Ultrasound

Figure 1: UMAP (Healy and McInnes, 2024) projections of InceptionV3 features of (pooled) ID and OOD image sets across four medical applications, demonstrating shifts between distributions of ID and OOD images. Within each application, the details of the specific ID and OOD image sets appear later in Section 4.1. Each point represents a single acquired medical image, projected from a 2048-dimensional feature space to 2 dimensions via UMAP.

cations spanning multiple imaging modalities and having significant diagnostic relevance, i.e., segmenting (i) the myocardium from cardiac magnetic resonance images (MRIs), (ii) the neuroretinal rim from retinal scans, (iii) the prostate from T2-weighted MRI, and (iv) the fetal head from ultrasound images; for which accurate segmentation are clinically vital. For instance, *myocardial segmentation* in short-axis cardiac MRI (Epstein, 2007; Petitjean and Dacher, 2011; Shaaf et al., 2022; Wang et al., 2015) is essential for estimating contractility and tissue strain, which aids in diagnosing infarction, ischemia, and ventricular dyssynchrony (Peng et al., 2016). In ophthalmology, optic disc and cup segmentation from retinal scans, i.e., the *neuroretinal rim*, enables the calculation of the cup-to-disc ratio (Fuchs and Duane, 1908; Lu, 2011), a key biomarker for monitoring glaucoma (Almazroa et al., 2015). Similarly, *prostate* segmentation in T2-weighted MRI is pivotal for the diagnosis, staging, and treatment planning of prostate cancer (Claus et al., 2004). Finally, *fetal head* segmentation in ultrasound images facilitates the measurement of geometric parameters to assess fetal growth and detect developmental anomalies (Zeng et al., 2022). Since each of these applications demands high-quality segmentation for patient care (Zeng et al., 2022; Almazroa et al., 2015; Peng et al., 2016), our work focuses on improving the quality of (OOD) segmentations to aid the subsequent clinical analysis.

We differentiate our method from a majority of existing segmentation-refinement methods that require the distribution-shifted data (which they call "OOD" data) during their training. Hence, such existing methods are inapplicable in our OOD setting where distribution-shifted data are unavailable during DNN training. Accordingly, we compare our method to only those existing approaches that train exclusively on ID data.

Conventional DNNs often require extensive training on large annotated datasets. Unlike natural images, which are abundant, and relatively easily acquired and annotated, medical images are typically scarce because they require expensive hardware for acquisition and specialized domain knowledge for accurate annotation. Hence, to mitigate this dependency on large training datasets, our DNN framework relies on a *lightweight architecture*. Our DNN trains on a *small but representative dataset* of only 100-200 pairs of medical images and their corresponding segmentation maps, using data-augmentation methods common in medical image analysis. This highlights our model's capability to learn from tiny training sets, an advantage in clinical settings.

This paper introduces VarDeepPCA, a novel lightweight variational DNN framework designed to restore/refine degraded segmentation maps by leveraging intrinsic geometric priors on the anatomical object of interest. Unlike existing approaches that require target-domain data or extensive pre-training, our VarDeepPCA explicitly learns a distribution of valid anatomical geometries using only small ID datasets. Our DNN framework learns the *principal modes of variation* in a class of segmentation maps, and models each segmentation map using a low-dimensional mixture-of-modes latent representation on a simplex. Theoretically, our novel *variational* learning framework leverages a reinterpretation of the softmax mapping to implicitly perform exact distribution modeling, thereby enabling computationally efficient, *sampling-free* learning and inference. This also enables VarDeepPCA to provide *uncertainty* estimates associated with its restored segmentation maps. We empirically validate our framework across 4 distinct clinical applications, using 14 publicly available datasets, involving segmentation of the myocardium, neuroretinal rim, prostate, and fetal head. Comparisons against 15 existing methods demonstrate that VarDeepPCA consistently restores segmentation maps produced by the existing methods on OOD data to (i) significantly improve anatomical plausibility of geometries and clinical utility of the segmentations, and (ii) significantly reduce errors, without needing any more training data than that used by existing methods.

The rest of the paper is organized as follows. Section 2 provides a comprehensive literature review in the

segmentation domain, covering the evolution of DNN architectures and loss functions, variational learning, uncertainty estimation, use of anatomical shape priors, and use of test-time-adaptive methods for generating robust segmentations. Section 3 describes our proposed method, the mathematical notations, the PCA-based latent representation, the variational interpretation of the softmax mapping, sampling-free variational learning, and demonstrates the use of our VarDeepPCA framework to improve existing segmenters on OOD images. Section 4 talks about the datasets, evaluation metrics, clinical utility of segmentation maps, baseline methods, implementation details, and the extensive results and discussions of both segmentation and uncertainty estimation on the four medical applications. Section 5 concludes the work by discussing the merits, some of the constraints in our framework, and future directions.

2. Related Works in Image Segmentation

DNNs have become the state of the art for many applications medical image segmentation. However, several challenges remain, particularly in achieving robustness to OOD data, ensuring anatomical plausibility, maintaining computational efficiency and providing uncertainty estimates, all while training on a small sample set. This section reviews existing methods in the aforementioned contexts.

2.1 Evolution of DNN Architectures and Loss Functions

Early DNN Methods for Image Segmentation. U-Net (Ronneberger et al., 2015) (here referred to as UNet) employs skip connections from the encoder to fuse context from the decoder with precise localization of anatomical objects. Attention U-Net (AttnUNet) (Oktay et al., 2018), a variant of UNet uses gating modules to focus on relevant regions during training, ResUNet (Zhang et al., 2018) uses residual units (He et al., 2016), which is then used as a backbone in ResUNet++ (Jha et al., 2019). DeepLabV3+ improves upon DeepLab (Chen et al., 2018b) architecture, enabling computationally efficient training. ResUNet++ uses squeeze-and-excitation (Hu et al., 2018) modules along with atrous spacial pyramid pooling (ASPP) modules introduced in DeepLabV3+ (Chen et al., 2018a) to create a parameter efficient segmentation architecture. However, such early DNNs typically lead to poor performance on OOD data because of their relatively lightweight architectures, straightforward loss functions, and the absence of prior modelling and pre-training (Yan et al., 2019; Torpmann-Hagen et al., 2022; Hendrycks et al., 2019; Pal et al., 2024).

Hybrid Loss Functions for Boundary Enhancement. Many DNN methods discussed earlier rely on a single loss term, e.g., binary cross-entropy (BCE) or soft-Dice (Galdran et al., 2022). Some later DNNs combine multiple loss terms (Qin et al., 2019; Sun et al., 2023; Kervadec et al., 2021) to focus on boundary regions. Some methods focus on the difference between the predicted boundary and the ground-truth segmentation, thereby, focusing on the loss around the segment boundary by employing difference-over-union (DoU) loss (Sun et al., 2023). Other methods use a boundary loss term by representing a non-symmetric L2 distance on the space of boundaries/contours as a regional integral. BASNet (Qin et al., 2019) uses a hybrid loss of BCE, structural-similarity (SSIM) (Wang et al., 2004; Zhao et al., 2017), and intersection over union (IoU) (Jadon, 2020). However, these DNNs are computationally heavier and often require pre-training on ImageNet-1K (Deng et al., 2009) dataset. Furthermore, these methods lack the knowledge of high-level segment characteristics related to the segmented object's geometry or topology (Jurdi et al., 2021; Gaikwad et al., 2023; Varma et al., 2023; Pal and Awate, 2024a; Gaikwad and Awate, 2024).

Generative Adversarial DNNs for Image Segmentation. Some methods use generative adversarial networks (GANs) (Goodfellow et al., 2014) towards medical image segmentation (Xun et al., 2022), primarily as a generative model for data augmentation (Andreini et al., 2020). SegAN (Xue et al., 2018) aims to improve segmentations by employing a critic to amplify the differences between the predicted/generated segmentations and the associated ground-truth segmentations. SegAN uses multi-scale L1 losses to capture both long-range and short-range spatial dependencies. SegAN aims to solve a min-max optimization problem, involving significant optimization challenges that are well known (Saxena and Cao, 2021).

Diffusion Models for Image Segmentation. Some recent approaches leverage diffusion processes for medical segmentation (Kazerouni et al., 2023). MedSegDiff (Wu et al., 2023) trains using denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020) with dynamic conditional encoding to mitigate noise. MedSegDiffV2 (Wu et al., 2024) extends MedSegDiff to segment multi-class objects via a transformer-based spectrum-space-former architecture. CIMD (Rahman et al., 2023) improves distribution modeling by using multiple expert annotations. DTAN (Zhao et al., 2024) incorporates text attention on diffusion for segmentation. However diffusion models for medical image segmentation demand substantial data and computational resources for optimization, while being more sensitive to hyperparameters (Wu et al., 2023; Ho et al., 2020). This makes OOD segmentation challenging for diffusion models (Zhang et al., 2025; Xie et al., 2025).

Transformers for Image Segmentation. Transformers, leveraging self-attention (Vaswani et al., 2017) mechanisms, have been applied to image segmentation (Thisanke et al., 2023; Li et al., 2024). Segmenter (Strudel et al., 2021) employs vision transformers (ViT) (Dosovitskiy et al.,

2021) and processes image patches. SegViT (Zhang et al., 2022) generates semantic segmentation masks through attention-to-mask modules. SegViT trains to translate learnable class tokens and spatial feature maps into segmentation maps. Some transformer architectures for medical image segmentation domain follow a U-shaped (Xiao et al., 2023) structure. DSTransUNet (Lin et al., 2021) improves upon TransUNet (Chen et al., 2024)[1] by employing hierarchical swin transformers to capture non-local and multiscale dependencies. It uses dual-scale encoders to extract coarse and fine-grained features across different semantic classes. These models are inherently bulky, and need pre-training on large datasets like ImageNet (Deng et al., 2009; Zheng et al., 2021; Pinto et al., 2021).

Foundational Models using Vision Transformers. Segment Anything Model (SAM) (Kirillov et al., 2023) leverages foundational models, modeling segmentation as a promptable task, enabling zero-shot transfer across a wide range of applications. While SAM demonstrates strong overall performance, it often misses fine structures, occasionally hallucinates small disconnected components, and may fail to produce crisp boundaries (Kirillov et al., 2023; Schiappa et al., 2024; Zhang et al., 2024). SAM often needs a lot of human interaction, through text prompts and image annotations, during inference. Furthermore, SAM was trained on a massive dataset of over one billion image-mask pairs, making training such models computationally expensive. MedSAM (Huang et al., 2024) extends SAM to medical images, but continues to share SAM's limitations.

State-Space Models for Image Segmentation. State-space models (SSMs) (Kalman, 1960), inspired by linear state-space equations in control theory, have recently emerged as an efficient alternative to transformers for modeling long-range dependencies, scaling linearly with sequence length. This includes pure SSM-based methods such as the Vision Mamba UNet (VM-UNet) (Ruan et al., 2024), which performs better than hybrid SSM-CNN models such as U-Mamba (Ma et al., 2024), SwinU-Mamba (Liu et al., 2024), and SegMamba (Xing et al., 2024). However, VM-UNet often struggles with low-contrast regions, is sensitive to image artifacts, and degrades its segmentation performance with higher image resolution.

2.2 Variational Learning and Uncertainty Estimation

Variational learning in DNNs models distributions in latent space during learning, e.g., the variational autoencoder (VAE) (Kingma and Welling, 2014), vector-quantized VAE (VQ-VAE) (van den Oord and Vinyals, 2017), and VQ-VAE2 (Razavi et al., 2019). Conditional VAEs (cVAEs) (Sohn et al., 2015) leverage latent variables to model conditional distributions on the output. Typical variational-learning methods introduce significant computational overhead, requiring expensive Monte-Carlo sampling (Mohamed et al., 2020) during both training and inference. Probabilistic UNet (ProbUNet) (Kohl et al., 2018) learns a distribution over segmentation maps for a given input image by combining UNet with cVAEs. Hierarchical ProbUNet (HierProbUNet) (Kohl et al., 2019) and PHiRec (Fischer et al., 2023) employ a cVAE with a hierarchical latent space. Probabilistic Hierarchical Segmentation (PHISeg) (Baumgartner et al., 2019) uses a VAE framework to model the conditional distribution of segmentation maps, for a given input image, when learning from multiple annotators at different spatial resolutions. However, many variational methods fail to provide robustness to OOD data (Gao et al., 2023b; Pal, 2021), because such models are not explicitly designed for domain shifts (Mehrtash et al., 2020; Gaikwad and Awate, 2021; Lennartz and Schultz, 2023).

Uncertainty-aware methods aim to output per-voxel uncertainty, often modeled as the standard deviation of a distribution on output segmentations, e.g., (Adiga et al., 2024). Some such approaches use Bayesian modeling and inference (Awate et al., 2019; Jena and Awate, 2019). Some methods quantify uncertainty estimates using normalized cross correlation (NCC) between the error maps and the uncertainty maps (Fischer et al., 2023). Another quantitative measure is the unified score (US; the higher the better) (Mehta et al., 2022) that combines area-under-curve (AUC) values of segmentation-performance measures (e.g., filtered true positive, filtered true negative, Dice similarity) with respect to segmentation-map thresholds. Some metrics focus on uncertainty calibration, e.g., adaptive calibration error (ACE) and thresholded ACE (TACE) (Nixon et al., 2019), extending expected calibration error (ECE) (Guo et al., 2017) using adaptive binning for robustness.

2.3 Anatomical Shape Priors

Some DNNs incorporate anatomical information by designing loss terms (Oktay et al., 2017; Jacob et al., 2024) that entail training on cross-domain data. However, in our (OOD) setting, data from the newer domain is unavailable during training. Some models require anatomical landmarks during inference (Gao et al., 2023a; Jacob et al., 2023), which must be provided either by an expert (who is unavailable our setting) or by a separate DNN (which itself would require training on the OOD data that is unavailable in our setting). Post-DAE (Larrazabal et al., 2020) enforces anatomical plausibility by projecting the predicted segmentation maps (degraded) generated on ID test data onto a manifold of valid shapes. Some existing methods leveraging anatomical priors, while not requiring OOD data, suffer from significant *computational and memory* limita-

1. The original article appeared on arXiv in 2021 (`https://arxiv.org/abs/2102.04306`)

tions, e.g., a recent VAE-based approach (Painchaud et al., 2020) learns valid cardiac shapes but requires empirically sampling millions of shapes by filtering them with a domain-specific correctness test, and storing (millions of) the valid ones. At inference it must perform an expensive search to find the nearest stored shape. Similarly, pointset-based shape priors (Shigwan et al., 2020) are often limited by Gaussian assumptions and expensive brute-force searches. Our framework avoids such limitations; it does not require any OOD data or annotations during training; by utilizing sampling-free variational learning, it avoids the need for simulation, storage, or any expensive inference-time search. Instead, it efficiently optimizes for the closest plausible segmentation using its DNN decoder via gradient descent on the space of valid segmentation-map geometries. Our model also provides uncertainty estimates, unlike existing methods using shape-priors for image segmentation.

2.4 Unsupervised Domain Adaptation

Existing DNN segmenters often fail in OOD domain because their models over-sensitive to the texture and spurious correlations within acquired medical images for generating the final predictions (Karani et al., 2021). To mitigate such effects, global intensity non-linear augmentation (GIN) and interventional pseudo-correlation augmentation (IPA) methods modify image appearance to break non-causal background links (Ouyang et al., 2022). Unsupervised domain adaptation (UDA) methods typically rely on aligning source data to the target data, without access to target-data labels. Source-free UDA (SFUDA) (Fang et al., 2024) methods, which are trained to adapt to unlabelled targets by using pre-trained models, are categorized into parameter-based white-box methods and output-based black-box methods. Methods such as DeY-Net learn features to encode the anatomy of medical objects for single domain generalization (SDG). Test-time adaptation (TTA) methods (Liang et al., 2025) adapts to target data using pre-trained models without access to source data and target-data labels, e.g., denoising TTA (DeTTA) (Wen et al., 2024). Adaptive Mutual Information (AdaMI) (Bateson et al., 2022) adapts to test data by fine-tuning models using label-free loss terms and anatomical class-ratio priors. Recently, SegCNN (Karani et al., 2021) adapts to OOD images using image-to-image normalization and DAEs. However, its performance degrades if the simulated noise during DAE training has a mismatch with the actual degradations in OOD images, or if the distribution shift renders the image-to-image normalization module ineffective (Karani et al., 2021).

2.5 Extensions Over Our Preliminary Work

This paper is a significant extension of our preliminary work in (Pal et al., 2025), advancing both its theoretical formulation and empirical analysis. The novel contributions of this extended work are as follows: (i) we propose a new mathematical formulation for the learning objective that explicitly models the output distribution's variance, unlike the formulation in (Pal et al., 2025) which disregarded this per-pixel uncertainty; (ii) we validate our method on more clinical applications: prostate segmentation in T2-weighted MRI and fetal head segmentation in ultrasound, incorporating a total of eight more datasets into our empirical analysis; (iii) we provide a new empirical sensitivity analysis for key hyperparameters, i.e., the size of the latent dimension K and the size of the training set S as discussed in Section 4.7, and show that our VarDeepPCA method performs robustly in all of the settings; (iv) we add seven new additional baselines for comparison, including the VMUNet (state-space model), the MedSAM foundational model specifically tuned for medical image segmentation and three new probabilistic baselines, i.e., Probabilistic UNet, Hierarchical Probabilistic UNet, and PHISeg, which produce both segmentation maps and their uncertainty estimates, along with the test-time-adaptation-based SegCNN method (with and without atlas), which is specifically designed to handle OOD data during test-time; (v) we add a comparative study for quantitative evaluation of the clinical utility of the uncertainty estimates produced by existing methods and our VarDeepPCA framework, using NCC, US, and TACE; (vi) we provide a more comprehensive discussion of the related works and methodology, and a more detailed presentation of the results including extensive qualitative and quantitative examples.

3. Methods

Our work is centered on a novel variational encoder-decoder DNN framework, namely *VarDeepPCA*, designed to learn a robust statistical model of geometry variability from ID segmentation maps. This learned model is then leveraged as a powerful prior to correct poor segmentations produced by existing DNNs on OOD data. The VarDeepPCA framework comprises: (i) a decoder that models the non-linear principal modes of anatomical geometry variation and their mixtures; (ii) a low-dimensional latent distribution that models these mixture proportions for a given segmentation map; (iii) an encoder that maps an input segmentation map to its corresponding latent distribution; and (iv) a *sampling-free* scheme for variational learning and inference. After learning this statistical geometry model, we employ a *fully automatic* correction process to improve the degraded segmentations produced by existing DNNs on OOD images.

3.1 Model Components and Mathematical Notation

Let X denote an acquired medical image containing an object of interest, and let Y be the associated expert-

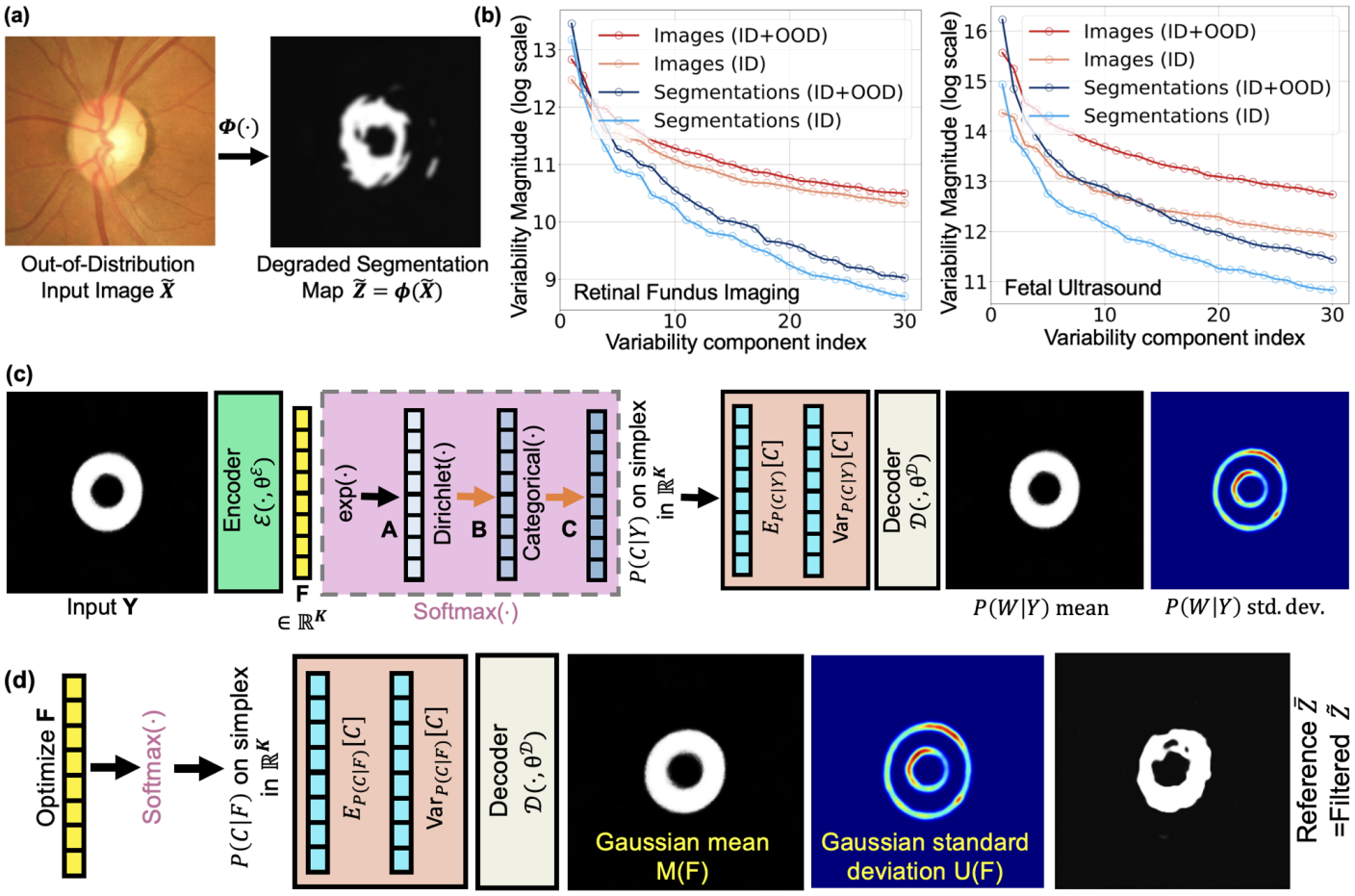


Figure 2: **(a)** Existing DNNs segment objects poorly on OOD images. **(b)** Principal (log) eigenvalues of covariance matrices of encodings in Inception DNN (Szegedy et al., 2016; Heusel et al., 2017) (in ID sets, as well as ID-union-OOD sets) show the variability across segmentation maps to be far lower (at least by an order of magnitude) than that across medical images (image intensities mapped to the range $[0, 1]$). **(c) Sampling-Free Variational Deep Learning of Principal Modes of Variation (VarDeepPCA) in Segmentation Maps.** Our VarDeepPCA models and learns principal modes of variation of (ID) segmentation maps. It models each segmentation map using a low-dimensional mixture-of-modes latent distribution on a simplex (Section 3.2) through a softmax mapping (Section 3.3). Our Bayesian interpretation of the softmax endows a variational model with (i) closed-form marginalization enabling sampling-free variational learning (Section 3.4) and (ii) per-pixel uncertainty estimates on test images (Section 3.4). **(d) VarDeepPCA Restores/Refines OOD-Image Segmentation Maps** by first "filtering" the degraded segmentation map, and then "projecting" the filtered segmentation map onto VarDeepPCA's learned principal modes of variation (Section 3.5).

annotated (binary/fuzzy) segmentation map. We differentiate this from W, which represents the (unknown) true anatomical segmentation. For a single given X, our framework is designed to handle cases with a single Y or multiple expert segmentations $\{Y_i\}$, each of which may differ from W. We consider an existing DNN segmenter $\Phi(\cdot)$, e.g., UNet, which is pre-trained on an ID dataset of (X, Y) pairs, with standard data-augmentation techniques. Our primary challenge arises when this segmenter is applied to an OOD image $\widetilde{X}$. By definition, $\widetilde{X}$ is drawn from a different distribution than the data used to train $\Phi(\cdot)$ stemming from domain shifts (Figure 1) because of the variation in imaging protocols across different hospitals and in imaging devices (e.g., different make/models, slight variations in calibration for identical instruments) (Liang et al., 2025). The domain shift may manifest as differences in the texture/appearance statistics (Karani et al., 2021). This distribution shift often causes the segmenter to produce a poor or anatomically implausible segmentation, $\widetilde{Z} := \Phi(\widetilde{X})$, as illustrated in Figure 2(a). Our goal is to correct $\widetilde{Z}$ without access to any OOD images $\widetilde{X}$ or their expert segmentations $\widetilde{Y}$ during the training of our VarDeepPCA model. In contrast, VarDeepPCA relies solely on the segmentation maps that were used to train $\Phi(\cdot)$, by constructing and employing priors on the anatomical-object geometry that remains largely invariant to these OOD variations, as illustrated in Figure 2(b). Indeed, the OOD description applies much more to the acquired medical images, resulting from variations in imaging pro-

tocols, devices, and reconstruction schemes across clinical sites, rather than variations in human anatomical geometry across the population. Figure 2(b) quantifies this variability in acquired medical images as well as the segmentation maps through the eigenvalues of the covariance matrix of latent Inception-DNN encodings.

3.2 DNN-based PCA Model on Segmentation Maps Using a Latent Representation on a Simplex

The VarDeepPCA framework employs a generic encoder-decoder DNN architecture to learn a statistical model of variability in segmentation maps for a class of objects (Figure 2(c)). We hypothesize that this variability can be captured by K *principal (non-linear) modes of variation*. In a latent space, we represent these discrete modes by a K-length *one-hot* random vector C, where $C = \mathbb{1}_k$ indicates the k-th mode, with a value of 1 at index k. The spatial-domain representation of these modes of variation is enabled by the mapping underlying our decoder. A typical segmentation map Y (in spatial domain) is not associated with a single mode (in latent space) but rather a *mixture* of these K principal modes. Therefore, we design VarDeepPCA to encode each Y into a *latent distribution* $P(C|Y)$. This distribution is represented by a K-dimensional vector of probabilities, $[P(C = \mathbb{1}_1|Y), \ldots, P(C = \mathbb{1}_K|Y)]$, which defines the association of Y with each of the K modes. As this vector's elements are non-negative and sum to 1, it is constrained to lie on a $(K-1)$-dimensional *simplex* in $\mathbb{R}^K$. To obtain this simplex representation, VarDeepPCA's encoder $\mathcal{E}(\cdot;\theta^{\mathcal{E}})$ first maps an input segmentation map Y to a K-dimensional *segmentation-feature* vector $F := \mathcal{E}(Y;\theta^{\mathcal{E}})$. We then apply the *softmax* function to F to produce the probability vector representing $P(C|Y)$. This use of the softmax is a critical design choice because it implicitly performs the necessary variational/distribution modeling (as described in Section 3.3) and, crucially enables *sampling-free* variational learning (as described in Section 3.4). Finally, VarDeepPCA's decoder $\mathcal{D}(\cdot;\theta^{\mathcal{D}})$ maps the latent simplex representation $P(C|Y)$ back to a distribution $P(W|Y)$ on the segmentation maps.

3.3 Reinterpreting Softmax Mapping in a Variational Setup

VarDeepPCA reinterprets the softmax function underlying the mapping $P(C|Y) \equiv \text{Softmax}(F)$ using Bayesian principles to illuminate the implicit variational modeling and distribution on the simplex in $\mathbb{R}^K$ (Figure 2(c)). We model $P(C|F)$ as the *posterior-predictive distribution* on C arising from (i) a *Categorical*-distribution *likelihood* $P(C|\cdot)$ on the modes of variation indicated by C, coupled with (ii) a *Dirichlet*-distribution (conjugate) *prior* $P(\cdot|F)$. Let the random vector A have elements $A_k := \exp(F_k) > 0$ for all $1 \leq k \leq K$, such that A parameterizes a Dirichlet distribution $\text{Dir}(B;A)$ of a hidden random vector B residing on the $(K-1)$-dimensional simplex. Since the mapping from $Y \to F \to A$ is deterministic, the following equivalence between posterior-predictive distributions holds: $P(C|Y = y) \equiv P(C|F = \mathcal{E}(y;\theta^{\mathcal{E}})) \equiv P(C|A = \exp(\mathcal{E}(y;\theta^{\mathcal{E}})))$. Consider a categorical distribution $\text{Cat}(C;B)$ on one-hot vectors C, which is parameterized by the hidden random vector B that is sampled from its *conjugate* (prior) distribution $\text{Dir}(B;A)$. The posterior-predictive distribution

$$P(C|A) = \int_b P(C|b)P(b|A)db \tag{1}$$

which equals

$$\int_b \text{Cat}(C;b)\text{Dir}(b;A)db \tag{2}$$

which equals

$$\int_b \Big(\prod_k (b_k)^{C_k}\Big)\Big(\frac{1}{\eta(A)}\prod_k (b_k)^{A_k-1}\Big)db, \tag{3}$$

where the normalizing constant for the Dirichlet distribution is $\eta(A) := \prod_k \Gamma(A_k)/\Gamma(\sum_k A_k)$, and $\Gamma(\cdot)$ denotes the Gamma function. This yields

$$P(C|A) = \frac{1}{\eta(A)}\int_b \prod_k (b_k)^{C_k+A_k-1}db = \frac{\eta(A+C)}{\eta(A)} \tag{4}$$

$$= \frac{\prod_k \Gamma(A_k+C_k)/\Gamma\left(\sum_k (A_k+C_k)\right)}{\prod_k \Gamma(A_k)/\Gamma(\sum_k A_k)}. \tag{5}$$

Now, consider a specific instance $C = \mathbb{1}_k$ (i.e., the k-th mode). Because C is a one-hot vector, $\sum_k C_k = 1$. Using the property $\Gamma(g+1) = g\Gamma(g)$, for gamma functions, we simplify the posterior-predictive distribution as

$$P(C = \mathbb{1}_k|A)$$

$$= \frac{\Gamma(A_k+1)\prod_{j\neq k}\Gamma(A_j)/\Gamma(\sum_j A_j+1)}{\prod_j \Gamma(A_j)/\Gamma(\sum_j A_j)} \tag{6}$$

$$= \frac{A_k\Gamma(A_k)\prod_{j\neq k}\Gamma(A_j)}{\prod_j \Gamma(A_j)} \cdot \frac{\Gamma(\sum_j A_j)}{(\sum_j A_j)\Gamma(\sum_j A_j)} \tag{7}$$

$$= \frac{A_k}{\sum_{j=1}^K A_j} = \frac{\exp(F_k)}{\sum_{j=1}^K \exp(F_j)}, \tag{8}$$

which is the k-th element of the $\text{Softmax}(F)$ vector. Thus, while the softmax mapping from F to the latent distribution $P(C|Y)$ is deterministic, it implicitly (i) subsumes variational modeling by defining the (prior) distribution $P(B|\exp(F)) \equiv P(B|A)$ and the (likelihood) distribution $P(C|B)$, and then (ii) marginalizes out the random variable B via Bayesian inference to produce the *analytically exact* posterior-predictive distribution $P(C|F)$ in *closed form*.

3.4 Sampling-Free Variational Learning

Let the training set of N segmentation maps be $\{Y_n\}_{n=1}^N$. For an input segmentation map Y, VarDeepPCA's internal low-dimensional representations (F and $P(C|F)$) are designed to model Y using only the top K modes of variation, thereby filtering out the remaining variation that arises from sources such as segmentation errors and discretization artifacts in Y. This is because the low-dimensional (K-dimensional) latent space in our autoencoder acts as a (well-studied) bottleneck with limited capacity (Laakom et al., 2024), which forces our autoencoder to model/represent mainly those dominant shapes/structures of the segmentation maps that were present in its training set of high-quality segmentation maps (Cho, 2013; Creswell and Bharath, 2018). For a given input Y, the variational model underlying VarDeepPCA produces a latent distribution $P(C|Y)$ by implicitly modeling the categorical distribution $P(C|B)$ and the Dirichlet distribution $P(B|Y=y) \equiv P(B|A=\exp(\mathcal{E}(y;\theta^{\mathcal{E}}))$. This enables VarDeepPCA to sample $c \sim P(C|Y)$ through the following procedure: (i) map input Y to $F \leftarrow \mathcal{E}(Y;\theta^{\mathcal{E}})$, (ii) map F to $A \leftarrow \exp(F)$, (iii) sample $b \sim \mathsf{Dir}(B;A)$, and (iv) sample $c \sim \mathsf{Cat}(C;b)$. The decoder then *outputs a distribution over segmentation maps* by mapping the latent distribution $P(C|Y)$ through the decoder $\mathcal{D}(\cdot)$. Specifically, the decoder maps each $c \sim P(C|Y)$ to a segmentation map, where we ensure that each per-pixel output lies within the range $[0,1]$ by incorporating a sigmoid layer as the final output layer of the decoder. For the (posterior-predictive) categorical distribution $P(C|Y)$, the mean and variance are available analytically in closed form:

$$C^{\mathsf{mean}} := \mathbb{E}_{P(C|Y;\theta^{\mathcal{E}})}[C] \tag{9}$$

$$= [P(C=\mathbb{1}_1|Y), \ldots, P(C=\mathbb{1}_K|Y)] \tag{10}$$

$$= \mathsf{Softmax}(\mathcal{E}(Y;\theta^{\mathcal{E}})) \tag{11}$$

(as per Section 3.3), and the k-th element of the variance is given by

$$C_k^{\mathsf{var}} := C_k^{\mathsf{mean}}(1 - C_k^{\mathsf{mean}}). \tag{12}$$

We model the decoder-output distribution by propagating the mean and variance of $P(C|Y)$ through the decoder, and approximating the output as a Gaussian distribution $\mathcal{N}(\cdot)$ characterized by a *mean segmentation map* $M := \mathcal{D}(C^{\mathsf{mean}};\theta^{\mathcal{D}})$ and a *variance map* V, which we describe next. Let $\mathcal{D}_i(\cdot)$ denote the decoder mapping to the i-th pixel. For pixel i in V, we model the variance V_i using (i) the variances C_k^{var} and (ii) the Jacobian of the decoder mapping $\mathcal{D}(L;\theta^{\mathcal{D}})$ (where L is a dummy variable) evaluated at C^{mean}. Thus,

$$M := \mathcal{D}(C^{\mathsf{mean}};\theta^{\mathcal{D}}), \text{ and} \tag{13}$$

$$V_i := \sum_{k=1}^{K} C_k^{\mathsf{var}} \left(\frac{\partial \mathcal{D}_i(L)}{\partial L_k}\Big|_{L:=C^{\mathsf{mean}}=\mathsf{Softmax}(\mathcal{E}(Y;\theta^{\mathcal{E}}))} \right)^2. \tag{14}$$

Our choice of modeling the output distribution as Gaussian stems from the Gaussian being the maximum-entropy (most general, in a sense) (Cover, 1999) distribution across all distributions constrained by a fixed mean M and a fixed variance V. From an alternative perspective, VarDeepPCA's output can be interpreted as (i) the representative segmentation M together with (ii) an underlying per-pixel *uncertainty* U (Figure 2(c)) given by the per-pixel square root of the values in V.

We formulate the variational learning objective to maximize, over parameters θ, the likelihood of the observed reference segmentation Y under the Gaussian distribution $\mathcal{N}(\cdot;M,V)$ output by the decoder. VarDeepPCA's variational learning formulation is therefore

$$\arg\max_{\theta} \prod_{n=1}^{N} \mathcal{N}(Y_n; M(Y_n;\theta), V(Y_n;\theta)) \equiv \tag{15}$$

$$\arg\min_{\theta} \sum_{n=1}^{N} \sum_{i=1}^{I} \frac{(Y_{ni} - M_i(Y_n;\theta))^2}{V_i(Y_n;\theta) + \epsilon} + \log(V_i(Y_n;\theta) + \epsilon), \tag{16}$$

where $M_i(\cdot)$ and $V_i(\cdot)$ denote, respectively, the values at the i-th pixel in the mean segmentation map M and the variance map V; $\epsilon > 0$ is a small regularization parameter for numerical stability. Thus, our VarDeepPCA learning formulation, despite explicitly modeling (i) a latent distribution $P(C|Y)$ and (ii) distributions $P(C|B)$ and $P(B|Y)$ implicitly within the softmax parameterization, eliminates the need for Monte Carlo sampling and the associated reparameterization that becomes necessary in typical variational deep networks (e.g., VAEs) due to the intractability of their underlying integrals.

3.5 VarDeepPCA to Improve Existing Segmenters on OOD Images

We propose a novel two-stage algorithm (Figure 2(d)) to leverage the learned VarDeepPCA model for restoring the poor segmentation maps $\widetilde{Z}$ produced by existing DNNs $\Phi(\cdot)$ on OOD images $\widetilde{X}$. In the first stage, we pass $\widetilde{Z}$ through the encoder-decoder of VarDeepPCA to "filter" out the non-principal components of variability from $\widetilde{Z}$, producing the "filtered" segmentation map

$$\overline{Z} := \mathcal{D}(\mathsf{Softmax}(\mathcal{E}(\widetilde{Z};\theta^{\mathcal{E}}));\theta^{\mathcal{D}}). \tag{17}$$

In the second stage, we explicitly "project" $\overline{Z}$ onto the learned space of principal modes of variation by (i) fixing $\overline{Z}$ as the output reference, (ii) optimizing the segmentation-feature vector in $\mathbb{R}^K$ as

$$F^* := \arg\max_F \mathcal{N}(\overline{Z}; M(F;\theta^{\mathcal{D}}), V(F;\theta^{\mathcal{D}})) \tag{18}$$

using gradient ascent, and (iii) obtaining the restored segmentation

$$M^* := \mathcal{D}(\text{Softmax}(F^*);\theta^{\mathcal{D}}) \tag{19}$$

with the associated per-pixel uncertainties given by $U_i^* := \sqrt{V_i^*}$ (Section 3.4).

In summary, VarDeepPCA introduces a principled variational framework that learns non-linear geometrical priors from ID data through a novel simplex-based latent representation, enabling sampling-free inference via softmax mappings that implicitly perform exact Bayesian marginalization. The framework's decoder outputs both a mean segmentation and per-pixel uncertainty estimates, providing interpretable measures of model confidence. We deploy this learned prior for the fully automatic correction of moderately degraded OOD segmentations by leveraging gradient-based projection onto the learned manifold of anatomically plausible geometries. The full procedure is detailed in Algorithm 1. We now proceed to empirically validate this framework on multiple clinical applications.

Algorithm 1 VarDeepPCA: Restoring OOD Segmentation Maps with Uncertainty Estimation.

Inputs:

- Degraded segmentation map $\widetilde{Z} := \Phi(\widetilde{X})$ from OOD image $\widetilde{X}$
- Trained VarDeepPCA model with encoder $\mathcal{E}(\cdot;\theta^{\mathcal{E}})$, decoder $\mathcal{D}(\cdot;\theta^{\mathcal{D}})$
- Projected gradient descent parameters: iterations T, learning rate η
- Small numerical constant $\epsilon > 0$

procedure RestorePoorSegmentation($\widetilde{Z}, \mathcal{E}, \mathcal{D}, T, \eta, \epsilon$)
  Stage 1: Filtering
  Extract features from degraded segmentation map: $F \leftarrow \mathcal{E}(\widetilde{Z};\theta^{\mathcal{E}})$
  Compute posterior-predictive distribution: $P(C|\widetilde{Z}) \leftarrow \text{softmax}(F)$
  Filter degraded segmentation map to give $\overline{Z} \leftarrow \mathcal{D}(P(C|\widetilde{Z});\theta^{\mathcal{D}})$
  Stage 2: Projection
  Initialize: $F^{(0)} \leftarrow F$ ▷ Start from extracted features
  for $t = 1$ **to** T **do**
    Compute latent-space distribution/mean: $L^{(t)} \leftarrow \text{softmax}(F^{(t-1)})$
    Decode to get mean map (spatial domain): $M^{(t)} \leftarrow \mathcal{D}(L^{(t)};\theta^{\mathcal{D}})$
    Compute latent-space variances: $C_k^{\text{var}} \leftarrow L_k^{(t)}(1 - L_k^{(t)}), \forall k$
    Compute variance map $V^{(t)}$ (spatial domain) at all pixels i as:
    $V_i^{(t)} \leftarrow \sum_{k=1}^K C_k^{\text{var}} \left(\frac{\partial \mathcal{D}_i(L)}{\partial L_k}|_{L=L^{(t)}}\right)^2$
    Compute Gaussian-based negative-log-likelihood loss:
    $\mathcal{L}^{(t)} \leftarrow \sum_{i=1}^I (\overline{Z}_i - M_i^{(t)})^2/(V_i^{(t)} + \epsilon) + \log(V_i^{(t)} + \epsilon)$
    Update using gradient descent: $F^{(t)} \leftarrow F^{(t-1)} - \eta \cdot \nabla_F \mathcal{L}^{(t)}$
  end for
  Set optimized features: $F^* \leftarrow F^{(T)}$
  Stage 3: Final Reconstruction and Uncertainty Quantification
  Compute projected latent-space distribution/mean: $L^* \leftarrow \text{softmax}(F^*)$
  Compute restored segmentation map (spatial domain): $M^* \leftarrow \mathcal{D}(L^*;\theta^{\mathcal{D}})$
  Compute latent-space variances: $C_k^{\text{var}^*} \leftarrow L_k^*(1 - L_k^*)$
  Compute uncertainty map U^* (spatial domain) at all pixels i as:
  $V_i^* \leftarrow \sum_{k=1}^K C_k^{\text{var}^*} \left(\frac{\partial \mathcal{D}_i(L)}{\partial L_k}|_{L=L^*}\right)^2$ and then $U_i^* \leftarrow \sqrt{V_i^*}$
  return M^*, U^*
end procedure

Outputs:

- Restored segmentation map (spatial domain) M^*
- Per-pixel uncertainty map (spatial domain) U^*

4. Results and Discussions

4.1 Datasets

We evaluate our framework across four distinct medical imaging applications: (i) segmenting the myocardium in MRI, (ii) segmenting the neuroretinal rim in retinal fundus images, (iii) segmenting the prostate in MRI, and (iv) segmenting the fetal head in ultrasound. These applications span diverse anatomical geometries, including genus-0 and genus-1 topologies. For each application, we train all models on a single dataset (ID training set) and test on one or more separate datasets (i.e., ID test set and OOD dataset). This cross-dataset evaluation scheme mimics a typical clinical scenario, testing robustness across domain shifts (Liang et al., 2025; Karani et al., 2021) caused by variations in imaging equipment, acquisition protocols, pathologies, etc. We pre-process images by cropping/padding and resampling image size to 256×256 pixels, applying data augmentation, and rescaling the intensities to the range $[0, 1]$. An overview of the datasets appears in Table 1.

Myocardium. We utilize three publicly available short-axis cardiac MRI datasets. The CAP dataset (Li et al., 2010; Kadish et al., 2009; Suinesiaputra et al., 2014), with 854 images, serves as the ID data, from which we use 150 images for training. The remaining data is split into a validation set (10% of the remainder, 70 images) and the rest (634 images) as an ID test set. For OOD evaluation, we employ the ACDC dataset (Bernard et al., 2018) (220 images) and the A-CMRI dataset (Andreopoulos and Tsotsos, 2008) (1722 images).

Neuroretinal Rim. For the neuroretinal rim segmentation task, we utilize three publicly available retinal fundus image datasets. The Magrabi dataset (Almazroa et al., 2018), with 833 images, serves as the ID data. From this dataset, we use 150 images for training, 63 for validation, and 620 for ID testing. For OOD evaluation, we employed the ORIGA dataset (Zhang et al., 2010) (637 images) and the G1020 dataset (Bajwa et al., 2020) (788 images).

Prostate. We use six publicly available T2-weighted MRI datasets from a multi-institutional collection for prostate segmentation (Liu et al., 2021), with a small number of images per institution. Within ID and OOD datasets, we desire to have a sufficient number of images and a comparable number of images. Thus, we design the ID and OOD datasets as follows: (i) the ID dataset combines BIDMC (Litjens et al., 2014) and BMC (Bloch et al., 2015; Barentsz et al., 2012), together having 381 images, (ii) the first OOD dataset combines HK (Litjens et al., 2014) and I2CVB (Lemaître et al., 2015), together having 366 images,

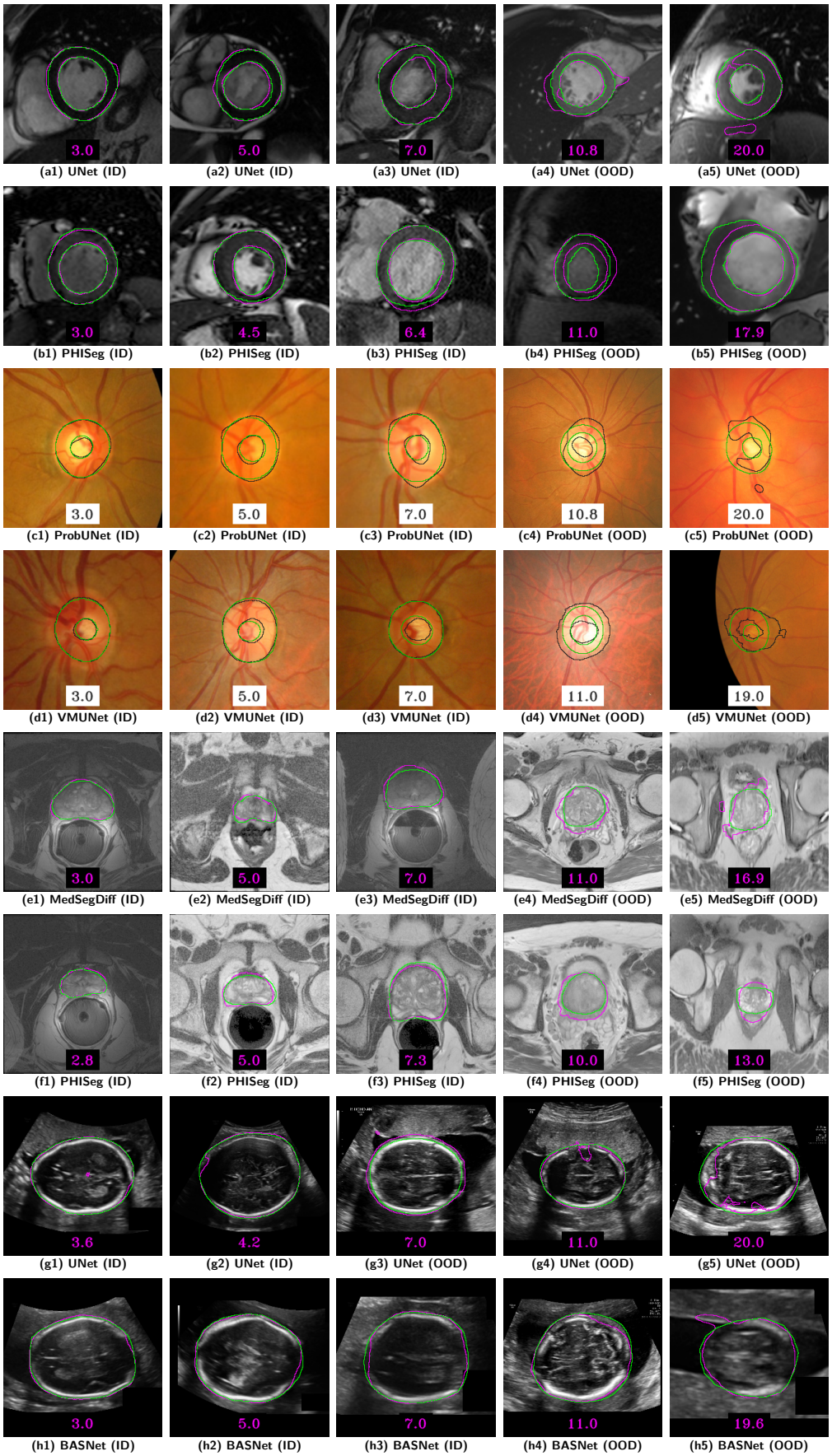


Figure 3: **Qualitative: Establishing Clinical Utility of Segmentation Maps.** Segmentation results across diverse baselines for: (i) myocardium on **(a1)-(a5)** UNet and **(b1)-(b5)** PHISeg; (ii) neuroretinal rim on **(c1)-(c5)** ProbUNet and **(d1)-(d5)** VMUNet; (iii) prostate on **(e1)-(e5)** MedSegDiff and **(f1)-(f5)** PHISeg; (iv) fetal head on **(g1)-(g5)** UNet and **(f1)-(f5)** BASNet. Ground-truth segmentation appears in green. The numbers show HD95 values.

Table 1: **Datasets.** Overview of the datasets used across the four medical imaging applications.

Application (Modality)	ID Dataset	ID Train	ID Val	ID Test	OOD Dataset 1 (Test Size)	OOD Dataset 2 (Test Size)
Myocardium (MRI)	CAP (Li et al., 2010; Kadish et al., 2009; Suinesiaputra et al., 2014)	150	70	634	ACDC (Bernard et al., 2018) (220 samples)	ACMRI (Andreopoulos and Tsotsos, 2008) (1722 samples)
Neuroretinal Rim (Fundus)	Magrabi (Almazroa et al., 2018)	150	63	620	ORIGA (Zhang et al., 2010) (637 samples)	G1020 (Bajwa et al., 2020) (788 samples)
Prostate (MRI)	BIDMC+BMC (Litjens et al., 2014; Bloch et al., 2015; Barentsz et al., 2012)	150	18	213	HK+I2CVB (Litjens et al., 2014; Lemaître et al., 2015) (366 samples)	RUNMC+UCL (Bloch et al., 2015; Barentsz et al., 2012; Litjens et al., 2014) (348 samples)
Fetal Head (Ultrasound)	HC18 (van den Heuvel et al., 2018)	150	68	666	FetalPlanes (Burgos-Artizzu et al., 2020) (1250 samples)	–

and (iii) the second OOD dataset combines RUNMC (Bloch et al., 2015; Barentsz et al., 2012) and UCL (Litjens et al., 2014), together having 348 images. From the ID dataset, we use 150 images for training, 18 images for validation, and 213 images for ID testing.

Fetal Head. We utilize two publicly available ultrasound datasets. The HC18 dataset (van den Heuvel et al., 2018), comprising 884 images, serves as the ID dataset. From this set, we use 150 images for training, 68 images for validation, and 666 images for ID testing. The FetalPlanes dataset (Burgos-Artizzu et al., 2020), containing 1250 images, is the OOD test set.

Curation of ID Training Set, ID Validation Set, ID Test Set. From the ID dataset, we aim to create an ID training subset that has a small size (say, 150) mimicking clinical scenarios, and that is representative of the diversity of anatomical shapes in the application. For this purpose, we propose the following strategy. First, in the ID dataset, for each binarized ground-truth segmentation map, we extract six features that collectively characterize the anatomical geometry (Alnaggar et al., 2024), i.e., region area, bounding-box area, extent, eccentricity, solidity, and orientation. Second, in this 6-dimensional feature space, we perform K-means clustering using 50 clusters. Third, we randomly select 3 segmentation maps from each of the 50 clusters, giving a total of 150 images (and their acquired medical images) in the ID training subset (used by all methods/baselines). From the remaining ID data, we use 10% as the ID validation set (to tune the hyperparameters for all models), and 90% as the ID test set. We train all the baselines, as well as VarDeepPCA, using the training dataset of these curated 150 image-mask pairs. We use the validation set to find the optimal latent-dimension size (i.e., K) for training our VarDeepPCA framework.

Data Augmentation. To enhance model generalization and mitigate overfitting, we employ data augmentation (Buslaev et al., 2020) during learning for VarDeepPCA as well as for all the baseline methods. The augmentation pipeline uses geometric and pixel-level transformations. Geometric augmentations, specifically horizontal flip, vertical flip, and affine transformations (comprising rotation, scaling, and translation), apply synchronously to medical images and their corresponding segmentation maps. Pixel-level augmentations include random brightness-contrast, random gamma, and blur.

4.2 Evaluation Metrics

Quantifying Segmentation Performance. Our evaluation prioritizes boundary-based metrics, which quantify distances between the predicted boundary points and the ground-truth boundary points. The Dice similarity coefficient (DSC) (Dice, 1945; Sørensen, 1948; Zijdenbos et al., 1994) is a popular metric for image segmentation, but it loses sensitivity when boundary-prediction errors are small relative to the size of the object (Seghier, 2024). Since our work focuses on precise object-boundary delineation, instead of localization, we primarily rely on metrics that are more sensitive to delineation errors (Nawaz et al., 2023). Our chosen metrics are the 95th-percentile Hausdorff distance (HD95) (Huttenlocher et al., 1993) and the average surface distance (ASD) (Yeghiazaryan and Voiculescu, 2018); both of these metrics are calculated in pixel distances, and lower values are better. These metrics are well-established for segmentation tasks in cardiac MRI (Bernard et al., 2018; Li et al., 2010), retinal images (Nawaz et al., 2023), prostate MRI (Langkilde et al., 2024; Liu et al., 2021), and fetal head ultrasound (Zeng et al., 2021; Nagabotu and Namburu, 2024). We also report DSC (higher is better) in percentage, for reference.

Table 2: **Model Sizes, Training Times, Inference Times, Hyperparameters.** The model size (number of parameters; in millions), training time (in minutes), and per-image inference time (in milliseconds) are the same across all four applications. MedSegDiff needed many more epochs during training, compared to other models. Learning rate (LR), weight decay (WD), batch size (BS), and loss functions are for training.

Models	Params (M)	Train. Time (mins)	Inf. Time (ms)	Train. Epochs	Key Hyperparameters, Loss Functions
UNet (Ronneberger et al., 2015)	31.03	6.8	9.9 ± 16.7	200	Adam: LR 1e-4, WD 5e-4; BS 64; Loss: SoftDice
AttnUNet (Oktay et al., 2018)	34.87	9.2	10.6 ± 27.3	200	Adam: LR 1e-4, WD 5e-4; BS 32; Loss: SoftDice
ResUNet++ (Jha et al., 2019)	4.06	7.3	9.1 ± 20.2	200	Adam: LR 1e-4, WD 5e-4; BS 64; Loss: SoftDice
DeepLabV3+ (Chen et al., 2018a)	59.33	5.6	10.1 ± 23.2	200	Adam: LR 1e-4, WD 5e-4; BS 64; Loss: SoftDice
BASNet (Qin et al., 2019)	87.06	21.4	18.2 ± 9.5	200	Adam: LR 1e-4, WD 5e-4; BS 24; Loss: BCE, SSIM, IoU
SegAN (Xue et al., 2018)	216.44	19.3	14.1 ± 21.3	200	Adam: LR 1e-4, WD 0, β 0.999; BS 24; Loss: SoftDice, Adv.
MedSegDiff (Wu et al., 2023, 2024)	129.40	2120.0	6.8e4 ± 150.1	10000	Adam: LR 1e-4, WD 0; EMA: 0.999; BS 14; Loss: MSE, Calib.
DSTransUNet (Lin et al., 2021)	171.44	17.8	61.2 ± 29.5	200	Adam: LR 1e-4, WD 5e-4; BS 24; Loss: Structure
VMUNet (Ruan et al., 2024)	44.27	21.1	20.7 ± 79.7	200	Adam: LR 1e-4, WD 5e-4; BS 8; Loss: SoftDice
MedSAM (Huang et al., 2024)	93.73	*NA*	1204.8 ± 30.3	*NA*	*NA*
PHISeg (Baumgartner et al., 2019)	99.21	12.9	183.7 ± 29.2	200	Adam: LR 1e-4; BS 16; Loss: Hierarchical KLD, CE
ProbUNet (Kohl et al., 2018)	5.0	7.6	21.0 ± 20.4	200	Adam: LR 1e-4, WD 0; Lat. Dim: 6; β: 10.0; BS 32; Loss: L2, ELBO
HierProbUNet (Kohl et al., 2019)	65.36	22.1	199.3 ± 40.0	200	Adam: LR 1e-4, WD 1e-5; Lat. Dim: 4; BS 8; Loss: BCE, GECO
SegCNN+TTA+DAE (+Atlas) (Karani et al., 2021)	2.25	8.08	3703.1 ± 272.9	200	Adam: Train LR 1e-4, WD 5e-4; BS 32; TTA LR 1e-3, steps 1000; Loss: SoftDice

Quantifying Uncertainty-Estimation Performance. We seek uncertainty estimates that correlate with, or are well-calibrated with respect to, actual errors in segmentation. For this purpose, we use three metrics (described in Section 2.2): (i) NCC (Fischer et al., 2023), (ii) US (Mehta et al., 2022), and (iii) TACE (Nixon et al., 2019).

4.3 Establishing Clinical Utility Based on Segmentation-Map HD95

We decide on the clinical utility of a given segmentation map generated by the DNN based on its HD95 value with respect to the ground-truth segmentation map, as follows. First, for the ID dataset in each application, we pool the segmentation results across all baselines and compute the histogram of the HD95 values. Subsequently, for each ID dataset, we analyze the histogram of HD95 values to find the typical range of HD95 values that can be considered as a clinically-acceptable performance (assuming most baselines work well on ID datasets). Through this quantitative analysis, we find that: (i) for the cardiac CAP dataset: HD95 $\leq$ 8 pixels accounted for around 92% of the segmentations; (ii) for the retinal MAGRABI dataset: HD95 $\leq$ 8 pixels accounted for around 72% of the segmentations; (iii) for the prostate BIDMC+BMC dataset: HD95 $\leq$ 8 pixels accounted for around 75% of the segmentations; (iv) for the fetal HC18 dataset: HD95 $\leq$ 8 pixels accounted for around 59% of the segmentations. Moreover, qualitative analysis in Figure 3 shows that HD95 values below 7-8 pixels typically indicate minor inconsistencies in the segmentation maps. Such inconsistencies may also arise from the variability across inter-expert and intra-expert annotations. On the other hand, HD95 values more than 11 typically indicate significant deviations from the ground truth. Thus, we consider a segmentation with a HD95 values $\leq$ 8 as having

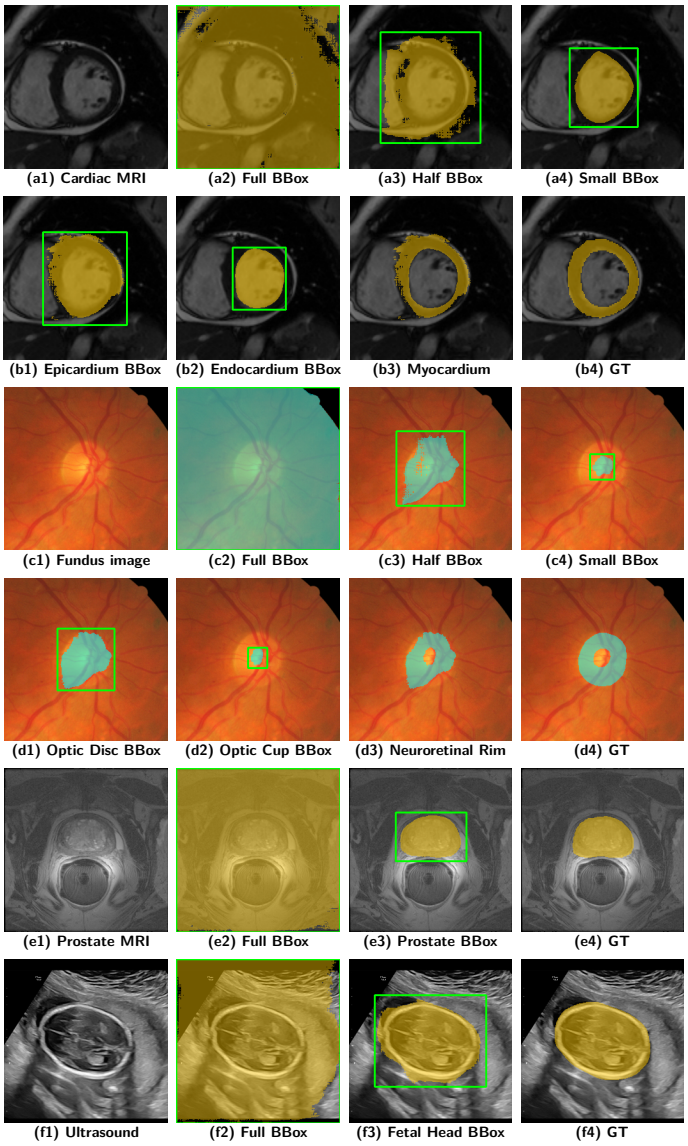


Figure 4: **MedSAM-based Segmentation using a Bounding Box (BBox) Prompt. (a1)–(b4)** Segmenting myocardium in cardiac MRI. **(c1)–(d4)** Segmenting neuroretinal rim in retinal fundus images. **(e1)–(e4)** Segmenting prostate in T2-weighted MRI. **(f1)–(f4)** Segmenting fetal head in ultrasound images.

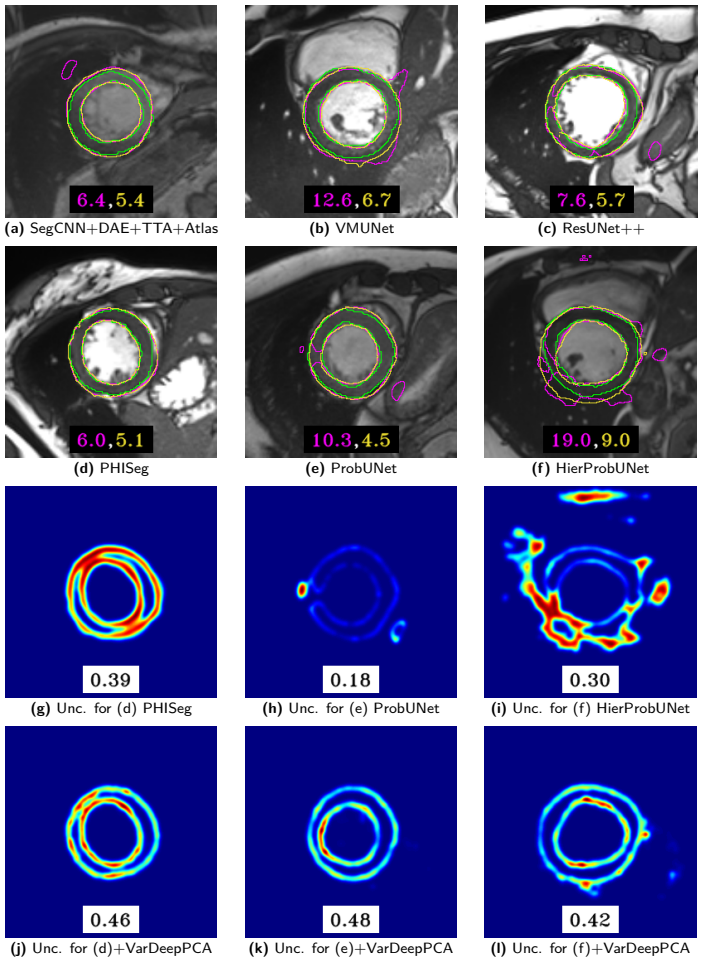


Figure 5: **Results–Qualitative: Myocardium Segmentation Restoration on ACDC (OOD) data. (a)–(c)** Results on images for the best non-variational baselines. **(d)–(f)** Results on images for the variational baselines. **(g)–(i)** Uncertainty maps produced using variational baselines. **(j)–(l)** Uncertainty maps produced using VarDeepPCA when plugged into the associated baselines (d)–(f). Color scheme in (a)–(f): Baseline; Baseline+VarDeepPCA (Ours); Ground Truth. HD95 numbers in (a)–(f) indicate that the examples were representative of the test set, because the HD95 values were close to the mean HD95 reported in Table 3. NCC numbers in (g)–(l) indicate that the examples were representative of the test set, because the NCC values were close to the mean of the NCC reported in Table 5.

clinical utility.

4.4 Baseline Methods

We compare our method against a comprehensive suite of baselines (denoted $\Phi(\cdot)$ earlier). As representative of standard encoder-decoder architectures (early DNN methods for medical image segmentation), we include UNet (Ronneberger et al., 2015), AttnUNet (Oktay et al., 2018), ResUNet++ (Jha et al., 2019), and DeepLabV3+ (Chen et al., 2018a). To represent hybrid loss functions, we use BASNet (Qin et al., 2019) that is designed for boundary-aware segmentation. To represent transformer architectures, we utilize DSTransUNet (Lin et al., 2021), a pre-trained model that has been shown to surpass TransUNet (Chen et al., 2024). Our generative baselines include the adversarial-based SegAN (Xue et al., 2018) and the diffusion-based MedSegDiff (Wu et al., 2023). We also use some very recent strategies, i.e., VMUNet (Ruan et al., 2024) representing state-space models, and MedSAM (Huang et al., 2024) representing foundation models. We evaluate MedSAM in a zero-shot setting using prompts (specific strategy described later), as it is a foundation model trained on over one million medical image-mask pairs.

Table 3: **Results–Quantitative: Myocardium Segmentation.** All models were trained on CAP (ID), and evaluated on ACDC and ACMRI (both OOD). For each method-dataset combination, we report mean (top row) and standard deviation (bottom row, in gray) for DSC(↑), HD95(↓), and ASD(↓). Augmenting each baseline with our VarDeepPCA consistently improves performance. **Bold-font** values in the columns indicate a statistically significant improvement of the Baseline+VarDeepPCA method over the underlying Baseline method, using a one-tailed paired-sample t-test ($p < 0.05$).

	CAP (ID)						**ACDC (OOD)**						**ACMRI (OOD)**					
	Baseline			Baseline + **VarDeepPCA**			Baseline			Baseline + **VarDeepPCA**			Baseline			Baseline + **VarDeepPCA**		
	DSC	HD95	ASD	DSC	HD95	ASD	DSC	HD95	ASD	DSC	HD95	ASD	DSC	HD95	ASD	DSC	HD95	ASD
UNet	89.6	6.2	2.3	**90.0**	**3.8**	**1.5**	73.5	28.6	7.7	**76.8**	**7.2**	**2.9**	75.2	26.0	7.9	**80.4**	**8.1**	**3.6**
	3.7	9.6	1.8	**3.4**	**1.3**	**0.5**	11.6	19.0	4.3	**9.9**	**3.7**	**1.2**	11.4	20.1	5.3	**8.1**	**3.5**	**1.8**
AttnUNet	90.9	3.6	1.5	90.9	3.5	**1.4**	78.4	15.3	4.3	**79.4**	**6.1**	**2.6**	75.5	17.7	5.6	**78.6**	**8.2**	**3.4**
	3.1	1.9	0.8	3.0	1.3	**0.5**	10.9	14.5	3.0	**9.1**	**3.0**	**1.1**	10.0	14.1	3.3	**8.5**	**3.3**	**1.5**
ResUNet++	89.1	4.2	1.6	**89.7**	**3.8**	**1.5**	77.1	9.3	2.9	77.4	**7.3**	**2.8**	78.2	11.8	4.1	**79.6**	**8.5**	**3.5**
	3.6	2.5	0.7	**3.5**	**1.3**	**0.5**	8.5	6.5	1.3	8.8	**3.5**	**1.1**	7.8	9.4	2.2	**7.4**	**3.1**	**1.3**
DeepLabV3+	88.4	4.4	1.8	**88.8**	**4.1**	**1.7**	70.1	13.6	5.1	70.5	**9.8**	**4.1**	69.9	11.8	4.5	**71.3**	**9.3**	**3.9**
	3.9	1.9	0.6	**3.9**	**1.2**	**0.5**	12.9	9.0	2.6	13.6	**4.4**	**1.7**	10.5	7.4	2.0	**10.4**	**2.8**	**1.1**
BASNet	91.3	3.2	1.4	91.4	3.1	1.3	80.2	10.0	3.5	80.2	**5.6**	**2.5**	81.1	9.1	3.4	81.2	**7.2**	**3.1**
	2.9	1.2	0.6	3.0	1.2	0.4	8.5	14.0	2.9	8.1	**1.8**	**0.9**	7.7	9.4	2.1	7.7	**2.8**	**1.3**
SegAN	91.6	3.3	1.3	91.7	3.2	1.3	72.3	11.7	4.3	**72.7**	**8.3**	**3.6**	71.0	10.9	4.0	**72.3**	**8.5**	**3.4**
	3.5	1.4	0.5	3.5	1.3	0.5	12.7	8.8	2.4	**12.8**	**3.2**	**1.3**	13.3	8.3	1.9	**12.7**	**3.1**	**1.2**
MedSegDiff	87.1	4.3	1.9	**88.1**	**4.2**	**1.9**	69.7	12.2	4.7	**71.0**	**8.9**	**4.1**	71.7	11.5	4.7	**73.0**	**9.5**	**4.3**
	7.3	1.6	0.7	**4.4**	**1.3**	**0.6**	10.0	8.7	2.3	**10.1**	**2.8**	**1.5**	12.1	8.4	2.3	**11.3**	**3.0**	**1.6**
DSTransUNet	91.5	3.5	1.3	**91.8**	**3.0**	**1.2**	77.6	11.6	3.8	**79.2**	**6.7**	**2.7**	81.0	8.3	3.2	**81.8**	**7.4**	3.2
	3.0	3.8	0.9	**2.8**	**1.2**	**0.5**	9.8	10.2	2.5	**9.0**	**3.5**	**1.2**	8.6	6.0	1.7	**7.9**	**3.7**	1.7
VMUNet	90.4	3.5	1.4	90.6	3.4	1.4	77.9	9.2	3.2	78.0	**6.9**	**2.8**	78.2	8.3	3.1	78.3	**7.9**	3.1
	2.9	1.2	0.5	2.9	1.2	0.5	10.4	9.2	2.4	10.2	**3.3**	**1.1**	10.4	4.8	1.4	10.5	**3.8**	1.3
MedSAM	62.9	10.4	4.1	**68.6**	**6.5**	**3.8**	60.0	9.5	3.1	**63.3**	**6.3**	**2.3**	71.3	8.8	3.3	**74.1**	**6.2**	**2.1**
	13.6	3.7	1.4	**13.1**	**3.1**	**1.5**	21.5	4.9	1.2	**20.6**	**4.1**	**0.2**	15.8	3.4	0.9	**14.0**	**3.1**	**0.3**
PHISeg	88.9	3.9	1.6	88.9	3.9	1.5	74.6	7.3	2.9	74.8	**6.7**	2.9	72.9	8.7	3.2	73.0	**8.3**	3.2
	4.2	1.6	0.6	4.2	1.4	0.6	14.2	4.2	1.2	14.3	**3.0**	1.1	11.1	3.6	1.0	10.9	**3.0**	1.0
ProbUNet	89.3	4.4	1.8	**90.1**	**3.6**	**1.5**	74.3	21.5	6.1	**78.1**	**6.4**	**2.6**	73.4	21.3	6.4	**77.9**	**7.6**	**3.2**
	3.9	3.2	1.0	**3.6**	**1.3**	**0.5**	10.8	17.0	3.9	**9.5**	**3.0**	**0.9**	10.4	15.0	3.7	**8.8**	**2.8**	**1.5**
HierProbUNet	86.7	6.2	2.3	**89.2**	**3.9**	**1.6**	60.4	39.1	12.0	**70.0**	**9.4**	**3.5**	71.0	25.7	7.8	**76.6**	**8.8**	**3.7**
	4.9	5.4	1.4	**4.0**	**1.4**	**0.5**	9.1	15.4	4.5	**8.2**	**2.9**	**1.2**	9.6	14.1	3.4	**8.6**	**3.0**	**1.5**
SegCNN+DAE +TTA	90.1	3.8	1.7	90.2	3.7	1.6	78.2	8.3	3.2	**79.7**	**5.9**	**2.2**	79.3	10.3	3.7	**80.4**	**7.8**	**2.7**
	3.2	1.4	0.5	3.2	1.4	0.4	10.1	7.5	2.2	**9.7**	**2.8**	**1.3**	9.8	9.4	2.8	**9.1**	**3.5**	**2.1**
SegCNN+DAE +TTA+Atlas	90.6	3.6	1.4	90.6	3.6	1.4	79.7	7.3	2.8	**80.6**	**5.8**	**2.1**	80.4	9.3	3.4	**81.9**	**7.5**	**2.6**
	3.2	1.3	0.4	3.2	1.3	0.4	9.8	7.4	1.9	**9.5**	**2.6**	**1.1**	9.0	8.9	2.5	**8.7**	**3.3**	**1.9**
Mean of Baselines	87.7	4.6	1.8	**88.6**	**3.8**	**1.6**	74.5	13.6	4.4	**75.8**	**7.2**	**3.0**	75.5	13.2	4.5	**77.5**	**8.0**	**3.3**
	8.9	4.0	1.2	**7.6**	**1.7**	**0.9**	12.2	13.6	3.3	**11.4**	**3.4**	**1.3**	11.4	12.2	3.1	**10.4**	**3.3**	**1.5**

Within the class of variational-learning methods (which also produce uncertainty estimates), we use ProbUNet (Kohl et al., 2018), HierProbUNet (Kohl et al., 2019), and PHISeg (Baumgartner et al., 2019). For each variational method, we generated 20 stochastic samples per input image at inference time; these samples lead to the mean segmentation map and the per-pixel standard deviation (uncertainty) map. We utilize the mean prediction generated by each variational method as the input for VarDeepPCA refinement. Within the class of TTA-based methods, we use the method by (Karani et al., 2021) that has (i) an image-to-image translator for normalization, which is adapted to the test image at test time, (ii) a UNet-based segmenter, and (iii) a denoising autoencoder (DAE) that removes minor inconsistencies in the segmentation masks. (Karani et al., 2021) uses jigsaw-based random-patch (Karani et al., 2021) replacement to train the DAE using 2D masks. We use two versions of the method by (Karani et al., 2021), i.e., SegCNN+TTA+DAE, and SegCNN+TTA+DAE+Atlas where the atlas constrains the predicted segmentation map to be close to typical ID segmentation maps.

Model Size, Training Time, and Inference Time.

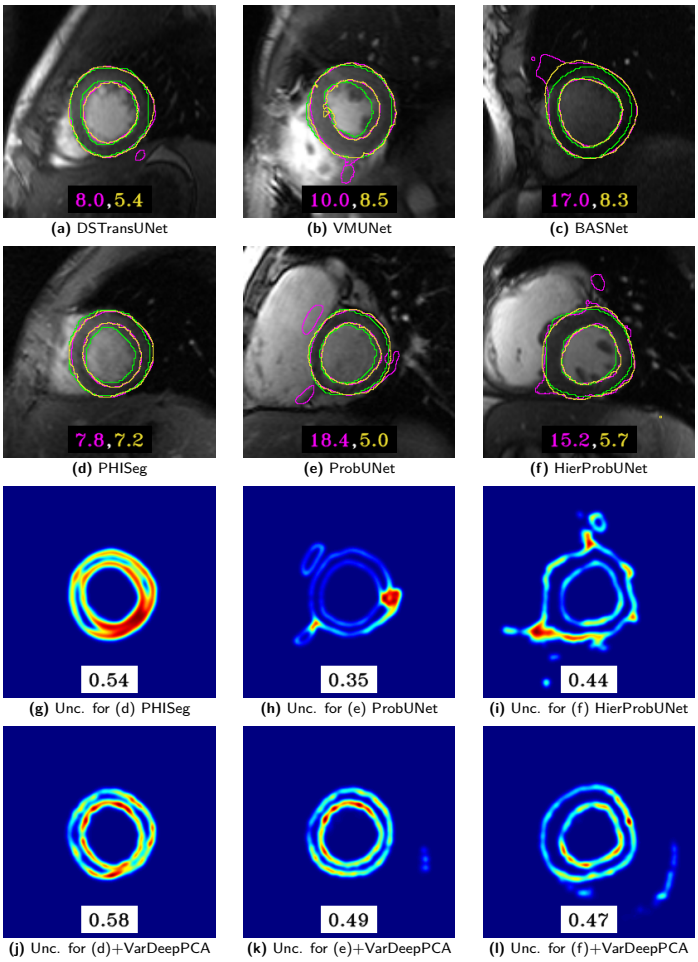


Figure 6: **Results–Qualitative: Myocardium Segmentation Restoration on ACMRI (OOD) data. (a)–(c)** Results on images for the best non-variational baselines. **(d)–(f)** Results on images for the variational baselines. **(g)–(i)** Uncertainty maps produced using variational baselines. **(j)–(l)** Uncertainty maps produced using VarDeepPCA when plugged into the associated baselines (d)–(f). Color scheme in (a)–(f): Baseline; Baseline+VarDeepPCA (Ours); Ground Truth. HD95 numbers in (a)–(f) indicate that the examples were representative of the test set, because the HD95 values were close to the mean HD95 reported in Table 3. NCC numbers in (g)–(l) indicate that the examples were representative of the test set, because the NCC values were close to the mean of the NCC reported in Table 5.

Table 2 gives the model size, training time, and inference time. All experiments use an NVIDIA RTX A6000 GPU. The training and inference times remain the same across applications because they use the same number of training images, same image size, and same number of epochs. Among the baselines, DeepLabV3+ exhibited the fastest training (5.6 min), likely due to its efficient atrous convolution modules (Chen et al., 2018a). ResUNet++ was the fastest during inference. The model in SegCNN+TTA+DAE had the smallest parameter count. Our VarDeepPCA plugin framework is lightweight. Its smallest configuration (latent dimension $K = 3$) comprises 1.02M parameters, which is approximately 50% of the smallest baseline (SegCNN+TTA+DAE). Its largest configuration (latent dimension $K = 16$) comprises 2.72M parameters; this contrasts sharply with the largest baseline (SegAN with 216.44M params). Because of VarDeepPCA's architectural simplicity, it has very small training times (under 2 minutes for all latent vectors from $K = 2$ to $K = 16$) and inference times (around 3.56 ms per image). In this way, the VarDeepPCA plugin typically leads to (very) small overheads in terms of model size and training/inference time.

MedSAM Bounding-Box Strategy. First, MedSAM's extensive training on over 1.5 million medical image-mask pairs (Huang et al., 2024) leads to a high risk of overlap between its training set and our test sets (giving it a potential undue advantage). Second, MedSAM leads to fundamental limitations in our applications: (i) MedSAM is highly sensitive to the placement and size of its bounding-box prompt, and (ii) MedSAM fails to generalize to anatomical topologies (e.g., ring shapes for myocardium and neuroretinal rim) that were absent in its training data, because its training was heavily biased towards genus-0 (Giri et al., 2021) objects. Consequently, a naively designed bounding-box prompt produces unusable results (Figure 4). Nevertheless, to establish a competitive baseline, we used an "oracle prompting" strategy that utilizes ground-truth information (of course, this is unfair to all other baselines) as follows: (i) to segment ring-like structures, we use MedSAM using a two-stage strategy that models the ring as a "subtraction" of an inner genus-0 segment from an outer genus-0 segment (Figure 4(a1)-(b4) for myocardium; Figure 4(c1)-(d4) for neuroretinal rim); (ii) for genus-0 structures, e.g., the prostate and fetal head, we use a single bounding box (Figure 4(e1)-(f4)); (iii) our bounding-box prompts use bounding boxes as expanded versions of the ground-truth bounding box, where the expansion adds a random margin of 10-30% of the box dimensions.

4.5 Implementation Details for Our VarDeepPCA

We demonstrate the benefits of VarDeepPCA using 14 publicly-available datasets and 15 existing DNN segmenters. VarDeepPCA's architecture uses: (i) an encoder $\mathcal{E}(\cdot;\theta^{\mathcal{E}})$ having a sequence of convolution and max-pooling layers that progressively increases the number of feature channels (32→64→128→256) while reducing spatial dimensions by 2× at each pooling stage (256→128→64→32→16), followed by a fully-connected bottleneck layer that projects the flattened 256×16×16 feature map to a low-dimensional latent vector $F \in 1 \times 1 \times \mathbb{R}^K$ (in this paper, K is tunable and set to the optimal value for each application), and (ii) a corresponding decoder $\mathcal{D}(\cdot;\theta^{\mathcal{D}})$ that applies a softmax acti-

Table 4: **Results–Quantitative: Neuroretinal Rim Segmentation.** All models were trained on MAGRABI (ID), and evaluated on G1020 and ORIGA (both OOD). For each method-dataset combination, we report mean (top row) and standard deviation (bottom row, in gray) for DSC(↑), HD95(↓), and ASD(↓). Augmenting each baseline with our VarDeepPCA consistently improves performance. **Bold-font** values in the columns indicate a statistically significant improvement of the Baseline+VarDeepPCA method over the underlying Baseline method, using a one-tailed paired-sample t-test ($p < 0.05$).

Models	MAGRABI (ID)						G1020 (OOD)						ORIGA (OOD)					
	Baseline			Baseline + **VarDeepPCA**			Baseline			Baseline + **VarDeepPCA**			Baseline			Baseline + **VarDeepPCA**		
	DSC	HD95	ASD	DSC	HD95	ASD	DSC	HD95	ASD	DSC	HD95	ASD	DSC	HD95	ASD	DSC	HD95	ASD
UNet	88.1	7.7	3.3	**89.4**	**6.0**	**2.8**	79.3	16.2	5.7	**82.6**	**7.3**	**3.4**	71.7	21.1	7.6	**76.3**	**7.6**	**4.3**
	4.7	6.3	1.9	**4.6**	**2.1**	**1.2**	10.8	19.3	5.9	**8.3**	**2.2**	**1.3**	10.2	20.5	6.2	**6.4**	**1.9**	**1.1**
AttnUNet	92.0	5.5	2.0	**92.9**	**4.9**	**1.8**	78.3	9.8	4.2	**82.2**	**7.7**	**3.4**	68.6	11.1	5.4	**69.3**	**9.3**	5.3
	4.0	2.4	0.8	**3.7**	**2.0**	**0.7**	9.5	7.1	1.9	**8.3**	**2.0**	**1.1**	7.3	4.2	1.3	**7.6**	**1.6**	1.1
ResUNet++	76.9	23.7	2.2	77.2	**7.1**	**1.9**	73.9	32.2	8.1	**76.6**	**10.8**	**5.3**	47.8	26.3	8.8	**55.7**	**10.4**	**5.7**
	21.9	12.2	1.8	21.5	**1.1**	**1.5**	10.6	16.9	7.3	**7.8**	**0.9**	**1.3**	25.4	15.9	6.3	**23.6**	**1.5**	**2.1**
DeepLabV3+	92.2	5.5	2.0	92.2	**5.3**	2.0	74.7	13.0	6.3	**76.3**	**9.2**	**5.5**	62.0	11.4	7.6	**63.3**	**8.6**	**6.5**
	3.7	2.5	0.9	3.6	**2.1**	0.8	7.4	8.4	2.4	**6.2**	**1.6**	**1.4**	7.4	4.0	1.3	**7.0**	**1.0**	**1.0**
BASNet	93.7	4.4	1.6	93.8	**4.3**	**1.6**	79.5	11.7	5.2	**80.5**	**8.0**	**4.4**	65.1	13.0	7.1	**67.6**	**10.1**	**6.5**
	3.0	1.9	0.7	2.9	**1.9**	**0.6**	7.1	10.4	2.8	**6.4**	**2.0**	**1.4**	6.8	8.9	2.2	**6.7**	**1.2**	**0.9**
SegAN	77.2	23.0	1.7	77.7	**7.2**	1.7	79.1	26.3	4.8	**83.0**	**10.7**	4.6	46.5	15.3	5.1	**53.9**	**10.5**	5.9
	23.0	10.9	0.7	22.9	**0.4**	1.0	9.0	3.2	1.6	**8.6**	**1.0**	1.5	19.2	7.8	1.8	**19.0**	**1.2**	1.3
MedSegDiff	92.5	5.2	1.9	**92.6**	**5.0**	**1.8**	77.8	9.8	5.3	78.0	**8.5**	**5.0**	62.9	10.8	7.0	**63.5**	**9.1**	7.0
	3.2	2.0	0.7	**3.2**	**2.0**	**0.7**	7.2	8.1	2.5	7.1	**1.9**	**1.6**	7.2	2.5	1.2	**7.2**	**1.2**	1.0
DSTransUNet	92.4	5.5	2.0	**93.1**	**4.7**	**1.7**	78.5	14.4	5.8	**80.7**	**7.9**	**4.1**	62.3	21.8	8.6	**68.2**	**9.6**	**6.0**
	3.3	4.3	1.0	**3.2**	**1.8**	**0.7**	8.6	16.5	5.3	**7.0**	**2.1**	**1.4**	8.1	15.4	3.6	**7.1**	**1.3**	**1.0**
VMUNet	93.0	4.8	1.8	**93.2**	**4.6**	**1.7**	79.1	10.0	4.8	**80.1**	**8.3**	**4.4**	61.8	12.7	7.2	62.0	**10.4**	**6.8**
	3.2	2.0	0.7	**3.3**	**2.0**	**0.7**	7.1	8.0	2.1	**6.8**	**1.9**	**1.4**	7.1	7.3	1.7	7.2	**1.2**	**1.0**
MedSAM	78.2	11.8	5.0	**84.6**	**7.8**	**3.3**	80.3	9.2	3.4	**85.1**	**6.8**	**2.9**	77.4	6.5	2.6	**81.1**	**5.5**	**2.4**
	7.6	3.2	1.5	**6.0**	**2.8**	**1.7**	8.5	4.0	1.0	**6.9**	**2.4**	**1.0**	11.6	2.1	0.7	**10.0**	**1.9**	**0.7**
PHISeg	92.3	5.0	1.9	**93.1**	**4.8**	**1.7**	74.6	10.7	4.2	**78.1**	**8.2**	**4.2**	63.6	10.4	6.3	63.5	**9.8**	6.4
	4.4	2.0	0.7	**3.2**	**1.9**	**0.6**	16.0	8.0	1.3	**9.7**	**1.9**	**1.3**	7.8	1.4	1.1	7.7	**1.3**	1.1
ProbUNet	82.9	10.1	4.3	**85.4**	**6.8**	**3.7**	66.0	21.8	7.8	**74.0**	**9.3**	**4.3**	62.5	30.7	10.9	**70.8**	**9.7**	**3.7**
	8.3	5.0	1.7	**7.4**	**2.4**	**1.6**	13.3	12.7	4.2	**11.5**	**1.8**	**1.2**	14.0	15.9	6.1	**13.6**	**1.6**	**0.9**
HierProbUNet	84.6	10.9	4.4	**88.7**	**6.3**	**3.0**	72.1	28.0	10.3	**80.8**	**8.6**	**4.0**	67.5	14.9	6.4	**72.0**	**9.1**	**4.4**
	6.1	8.1	3.0	**4.7**	**2.3**	**1.1**	11.1	26.8	9.8	**8.5**	**2.1**	**1.4**	8.7	12.6	5.2	**8.1**	**1.8**	**1.2**
SegCNN+DAE +TTA	89.3	6.8	2.9	**90.7**	**6.3**	**2.7**	79.6	9.7	4.9	**80.8**	**8.1**	**4.3**	67.9	10.5	6.6	**68.5**	**9.5**	**5.1**
	4.3	2.6	1.3	**3.6**	**2.1**	**1.7**	8.5	7.3	3.1	**7.5**	**2.6**	**1.6**	8.3	4.9	1.9	**7.9**	**2.1**	**1.8**
SegCNN+DAE +TTA+Atlas	90.3	6.2	2.5	**91.2**	**5.7**	**2.1**	80.1	8.9	4.2	**81.1**	**7.7**	**3.8**	69.1	9.9	5.4	**70.2**	**9.0**	**4.9**
	4.0	2.4	1.0	**3.9**	**2.1**	**1.0**	7.3	6.0	2.0	**7.1**	**2.0**	**1.2**	7.6	4.1	1.5	**7.5**	**1.6**	**1.1**
Mean of Baselines	89.5	6.8	2.7	**90.9**	**5.6**	**2.3**	77.6	12.6	5.2	**80.4**	**8.1**	**4.1**	67.5	14.1	6.5	**70.2**	**8.8**	**5.0**
	6.6	4.5	1.7	**5.2**	**2.3**	**1.3**	10.3	12.5	4.0	**8.3**	**2.2**	**1.5**	10.5	12.2	3.9	**10.1**	**2.2**	**1.8**

vation to F before using a fully-connected expansion layer and transpose-convolutions for upsampling back to the original 256×256 resolution. Training uses Adam (Kingma and Ba, 2015), a batch normalization after each convolution layer, and four independent runs from which we choose the best model. This architecture achieves dimensionality reduction by compressing high-dimensional spatial features into a K-dimensional representation, enabling efficient latent space analysis while learning to preserve reconstruction quality through the decoder pathway. The specific model configurations were tailored to each application: $K = 16$ (2.72M parameters) for the myocardium; $K = 8$ (1.67M parameters) for both the neuroretinal rim and prostate; and $K = 3$ (1.02M parameters) for the fetal head. All implementations use PyTorch (Paszke et al., 2019). In the interest of reproducibility, the code and models used in this study will be made publicly available after publication.

4.6 Results: Quantitative and Qualitative

A comprehensive analysis across all four application domains reveals degraded OOD performance for many early-DNN

Table 5: **Results–Quantitative: Myocardium – Measuring Calibration between Per-Pixel Segmentation Uncertainty and Per-Pixel Segmentation Error.** All models were trained on CAP (ID), and evaluated on ACDC and ACMRI (both OOD). For each method-dataset combination, we report the mean (top row) and standard deviation (bottom row; in gray) for NCC (↑), US (↑), and TACE (↓) metrics. Augmenting each baseline with our VarDeepPCA shows better calibration. **Bold-font** values in the columns indicate a statistically significant improvement of the Baseline+VarDeepPCA method over the underlying Baseline method, using a one-tailed paired-sample t-test ($p < 0.05$).

	CAP (ID)						**ACDC (OOD)**						**ACMRI (OOD)**					
	Baseline			Baseline + **VarDeepPCA**			Baseline			Baseline + **VarDeepPCA**			Baseline			Baseline + **VarDeepPCA**		
	NCC	US	TACE	NCC	US	TACE	NCC	US	TACE	NCC	US	TACE	NCC	US	TACE	NCC	US	TACE
PHISeg	0.53	0.83	0.23	**0.58**	**0.85**	**0.18**	0.41	0.73	0.35	**0.48**	**0.88**	**0.16**	0.52	0.75	0.32	**0.56**	**0.85**	**0.10**
	0.04	0.03	0.05	**0.03**	**0.02**	**0.03**	0.05	0.08	0.08	**0.05**	**0.07**	**0.03**	0.04	0.07	0.06	**0.05**	**0.06**	**0.02**
ProbUNet	0.34	0.75	0.29	**0.50**	**0.91**	**0.19**	0.23	0.63	0.31	**0.46**	**0.89**	**0.21**	0.33	0.73	0.34	**0.45**	**0.87**	**0.12**
	0.06	0.01	0.06	**0.03**	**0.01**	**0.03**	0.05	0.04	0.12	**0.04**	**0.04**	**0.03**	0.08	0.04	0.10	**0.04**	**0.05**	**0.02**
HPUNet	0.45	0.73	0.18	**0.50**	**0.85**	**0.10**	0.34	0.61	0.49	**0.43**	**0.75**	**0.26**	0.42	0.57	0.31	**0.47**	**0.82**	**0.19**
	0.03	0.02	0.07	**0.03**	**0.02**	**0.03**	0.04	0.04	0.07	**0.05**	**0.04**	**0.02**	0.05	0.05	0.09	**0.04**	**0.04**	**0.02**

methods and some variational methods. While the variational model PHISeg demonstrated considerable robustness to OOD data and provided meaningful uncertainty maps, other large architectures, e.g., BASNet, MedSegDiff, and DSTransUNet, showed only occasional robustness depending on the specific application and the specific OOD dataset. We find that SegCNN+DAE+TTA+Atlas shows robust segmentation performance on OOD data in some applications. Our VarDeepPCA framework consistently refined the results, with often significantly improving the results produced by the baseline methods. In datasets where baseline performance had poor accuracy and poor precision (exhibiting high variance) in terms of HD95, VarDeepPCA not only improved the accuracy but also the precision. The reduction in variance may be attributed to VarDeepPCA's ability to learn the manifold/distribution of anatomically valid geometries. For qualitative analysis, we show representative examples for which the HD95 value comes close to the mean HD95 value across the entire test set.

Myocardium Segmentation. On the CAP (ID) test set, most baseline models perform well (Table 3), achieving a mean DSC $\geq$ 86% and HD95 $\leq$ 6.5; this implies that the models fitted well (avoiding overfitting) to the training data. As anticipated, when the same models are applied to OOD datasets (ACDC and ACMRI), we observe a significant degradation in performance across all metrics; this confirms the baseline models' lack of robustness to domain shifts. Our primary finding is that plugging in our VarDeepPCA framework consistently and significantly lowers the HD95 and ASD values across all baselines on these OOD datasets. For these experiments, we use a latent dimension of $K = 16$ for VarDeepPCA, as justified in our sensitivity

Table 6: **Results–Quantitative: Neuroretinal Rim – Measuring Calibration between Per-Pixel Segmentation Uncertainty and Per-Pixel Segmentation Error.** All models were trained on MAGRABI (ID), and evaluated on G1020 and ORIGA (both OOD). For each method-dataset combination, we report the mean (top row) and standard deviation (bottom row; in gray) for NCC (↑), US (↑), and TACE (↓) metrics. Augmenting each baseline with our VarDeepPCA shows better calibration. **Bold-font** values in the columns indicate a statistically significant improvement of the Baseline+VarDeepPCA method over the underlying Baseline method, using a one-tailed paired-sample t-test ($p < 0.05$).

	MAGRABI (ID)						**G1020 (OOD)**						**ORIGA (OOD)**					
	Baseline			Baseline + **VarDeepPCA**			Baseline			Baseline + **VarDeepPCA**			Baseline			Baseline + **VarDeepPCA**		
	NCC	US	TACE	NCC	US	TACE	NCC	US	TACE	NCC	US	TACE	NCC	US	TACE	NCC	US	TACE
PHISeg	0.56	0.93	0.21	**0.58**	**0.95**	**0.12**	0.48	0.82	0.32	**0.57**	**0.87**	**0.13**	0.54	0.82	0.48	**0.56**	**0.86**	**0.18**
	0.05	0.03	0.04	**0.04**	**0.02**	**0.01**	0.07	0.11	0.06	**0.05**	**0.06**	**0.01**	0.05	0.04	0.06	**0.04**	**0.03**	**0.01**
ProbUNet	0.42	0.91	0.27	**0.49**	**0.94**	**0.12**	0.37	0.85	0.39	**0.43**	**0.88**	**0.13**	0.33	0.86	0.55	**0.39**	**0.88**	**0.16**
	0.07	0.04	0.08	**0.06**	**0.03**	**0.02**	0.08	0.06	0.12	**0.07**	**0.06**	**0.01**	0.08	0.07	0.11	**0.05**	**0.06**	**0.01**
HPUNet	0.45	0.92	0.19	**0.53**	**0.95**	**0.10**	0.44	0.88	0.40	**0.49**	**0.92**	**0.13**	0.43	0.88	0.45	**0.53**	**0.93**	**0.12**
	0.04	0.03	0.09	**0.05**	**0.02**	**0.01**	0.07	0.06	0.12	**0.07**	**0.04**	**0.01**	0.05	0.04	0.09	**0.05**	**0.04**	**0.01**

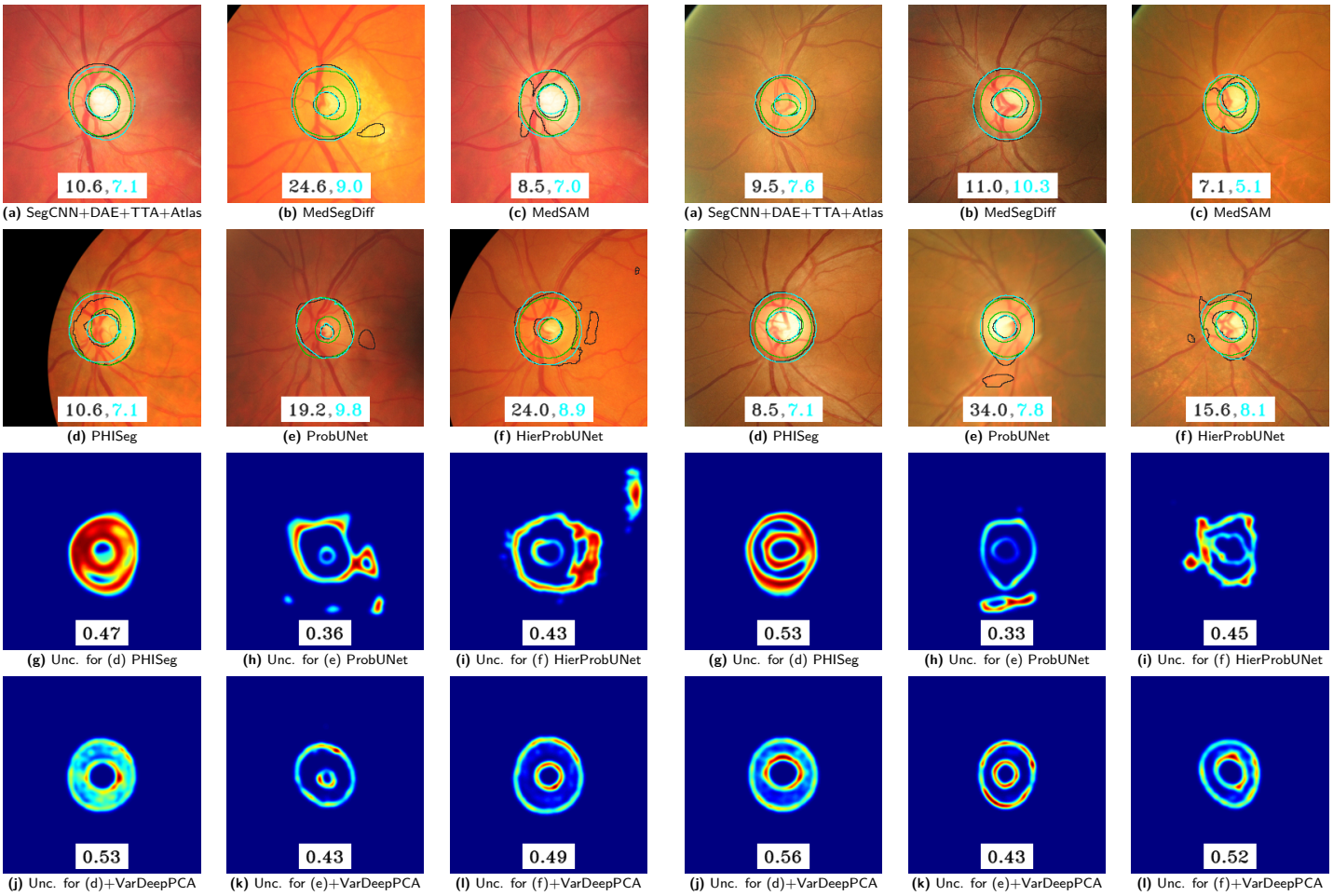


Figure 7: **Results–Qualitative: Neuroretinal Rim Segmentation Restoration on G1020 (OOD) data. (a)–(c)** Results on images for the best non-variational baselines. **(d)–(f)** Results on images for the variational baselines. **(g)–(i)** Uncertainty maps produced using variational baselines. **(j)–(l)** Uncertainty maps produced using VarDeepPCA when plugged into the associated baselines (d)–(f). Color scheme in (a)–(f): Baseline; Baseline+VarDeepPCA (Ours); Ground Truth. HD95 numbers in (a)–(f) indicate that the examples were representative of the test set, because the HD95 values were close to the mean HD95 reported in Table 4. NCC numbers in (g)–(l) indicate that the examples were representative of the test set, because the NCC values were close to the mean of the NCC reported in Table 6.

Figure 8: **Results–Qualitative: Neuroretinal Rim Segmentation Restoration on ORIGA (OOD) data. (a)–(c)** Results on images for the best non-variational baselines. **(d)–(f)** Results on images for the variational baselines. **(g)–(i)** Uncertainty maps produced using variational baselines. **(j)–(l)** Uncertainty maps produced using VarDeepPCA when plugged into the associated baselines (d)–(f). Color scheme in (a)–(f): Baseline; Baseline+VarDeepPCA (Ours); Ground Truth. HD95 numbers in (a)–(f) indicate that the examples were representative of the test set, because the HD95 values were close to the mean HD95 reported in Table 4. NCC numbers in (g)–(l) indicate that the examples were representative of the test set, because the NCC values were close to the mean of the NCC reported in Table 6.

analysis in Section 4.7. This demonstrates that VarDeepPCA successfully filters the degraded segmentation maps and projects them onto the learned manifold of anatomically plausible geometries. Notably, VarDeepPCA also improves the boundary metrics (HD95, ASD) even on the ID test set. This suggests that our model is not just correcting for large OOD shifts but also filtering minor anatomical inaccuracies present in the baselines' ID outputs. While the corresponding gains in DSC are more modest (since this metric is less sensitive to the boundary errors that are the focus of our work), we argue that the substantial improvement in boundary-based metrics brings the segmentations closer to the ground truth, achieving a level of accuracy more suitable for clinical applications. Indeed, for the OOD datasets of ACDC and ACMRI, the mean HD95 values after employing the VarDeepPCA plugin reduce from an average (across all methods; before VarDeepPCA) within 13.2-13.6 to an average (across all methods; after VarDeep-

PCA) within 7.2-8, which is far more clinically acceptable as per our analysis in Section 4.3. We also present the quantitative results for the uncertainty calibration metrics, i.e., the NCC, US, and TACE scores in Table 5. For the PHISeg, ProbUNet, and HPUNet baselines, incorporating the VarDeepPCA plugin leads to a significant improvement of these scores in both ID and OOD datasets.

For qualitative analysis, we selected the three strongest-performing non-variational baselines for each OOD dataset: MedSAM, VMUNet, and ResUNet++ for ACDC (OOD); DSTransUNet, VMUNet, and BASNet for ACMRI (OOD). As shown in Figure 5 and Figure 6, these baselines often exhibit common OOD failure modes, such as disconnected blob-like structures, potentially arising from bias-field and other artifacts introduced by different acquisition equipment. We also analyzed all variational baselines. We found that PHISeg, a large variational model with 99.21M parameters (Table 8), shows considerable robustness on both ID and OOD data, which we attribute to its complex multi-scale architecture. Conversely, ProbUNet and HierProbUNet architectures were more sensitive to the domain shift, resulting in segmentations with highly uncertain and erroneous boundaries. However, ProbUNet and HierProbUNet performed reasonably well on the CAP (ID) dataset. Figure 5 and Figure 6 demonstrate that our VarDeepPCA not only restores segmentation maps to make them more consistent with the underlying anatomy, but it also produces uncertainty maps that show significantly more localization of uncertainty (aligned with the geometry of the anatomical object of interest) and lower overall uncertainty compared to the variational baselines.

Neuroretinal Rim Segmentation. We trained all models on the MAGRABI (ID) dataset, and evaluated their performance on the MAGRABI (ID) test, G1020 (OOD), and ORIGA (OOD) datasets. On the ID test set, most baselines yielded competitive HD95 scores (Table 4). Exceptions included ResUNet++ that is likely constrained by its limited architectural capacity, and SegAN that likely suffered from the known instability of adversarial optimization. On the OOD datasets, we observed a substantial drop in DSC values. Notably, this DSC degradation was uniform across most baselines. However, the boundary-based metrics (HD95 and ASD) showed different amounts of performance degradations across baselines. In this context, our VarDeepPCA helped improve the baselines segmentation maps, leading to a mean HD95 $\leq$9 on G1020 and $\leq$10 on ORIGA, significantly minimizing boundary errors compared to the baselines. For this application, we configured the VarDeepPCA architecture with a latent dimension of $K = 8$ (refer Section 4.7 for sensitivity analysis). Indeed, for the OOD datasets of G1020 and ORIGA, the mean HD95 values after employing the VarDeepPCA plugin reduce from an average (across all methods; before VarDeepPCA) within

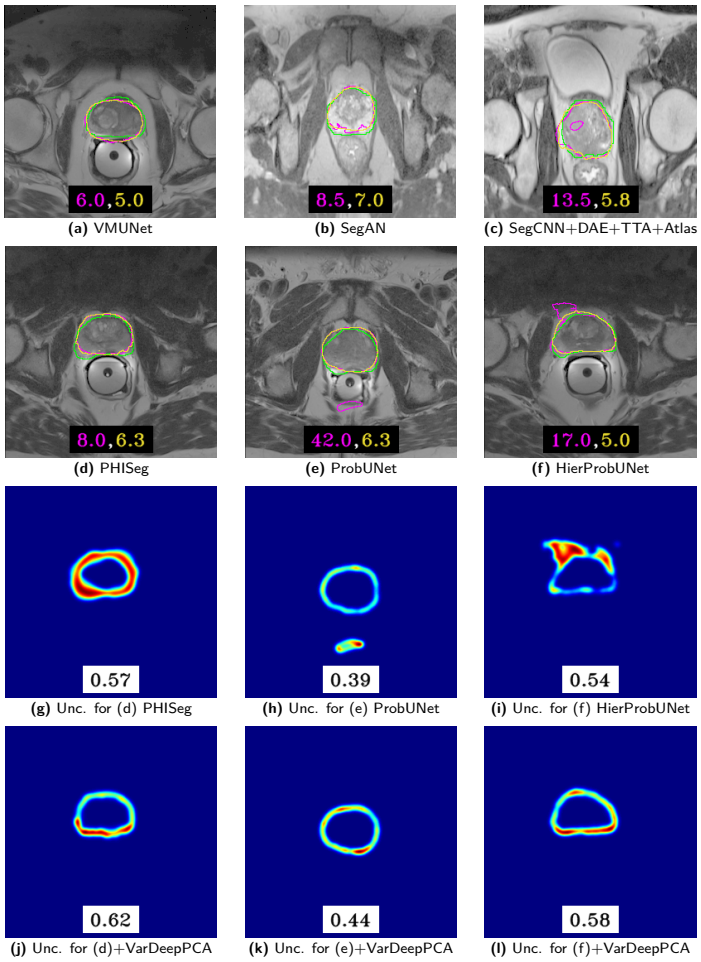


Figure 9: **Results–Qualitative: Prostate Segmentation Restoration on HK+I2CVB (OOD) data. (a)–(c)** Results on images for the best non-variational baselines. **(d)–(f)** Results on images for the variational baselines. **(g)–(i)** Uncertainty maps produced using variational baselines. **(j)–(l)** Uncertainty maps produced using VarDeepPCA when plugged into the associated baselines (d)–(f). Color scheme in (a)–(f): Baseline; Baseline+VarDeepPCA (Ours); Ground Truth. HD95 numbers in (a)–(f) indicate that the examples were representative of the test set, because the HD95 values were close to the mean HD95 reported in Table 7. NCC numbers in (g)–(l) indicate that the examples were representative of the test set, because the NCC values were close to the mean of the NCC reported in Table 9.

12.6-14.1 to an average (across all methods; after VarDeepPCA) within 8.1-8.8, which is far more clinically acceptable as per our analysis in Section 4.3. We also present the quantitative results for the uncertainty calibration metrics, i.e., the NCC, US, and TACE scores in Table 6. For the PHISeg, ProbUNet, and HPUNet baselines, incorporating the VarDeepPCA plugin leads to a significant improvement of these scores in both ID and OOD datasets.

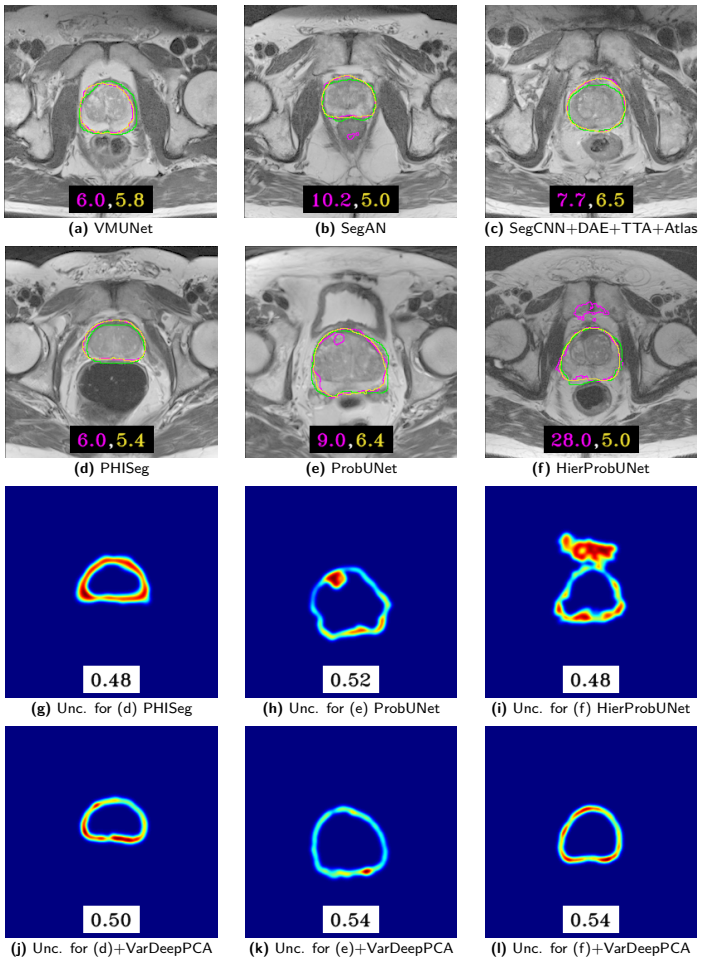
We compare with the strongest non-variational base-

Table 7: **Results–Quantitative: Prostate Segmentation.** All models were trained on BIDMC+BMC (ID), and evaluated on HK+I2CVB and RUNMC+UCL (both OOD). For each method-dataset combination, we report mean (top row) and standard deviation (bottom row, in gray) for DSC(↑), HD95(↓), and ASD(↓). Augmenting each baseline with our VarDeepPCA consistently improves performance. **Bold-font** values in the columns indicate a statistically significant improvement of the Baseline+VarDeepPCA method over the underlying Baseline method, using a one-tailed paired-sample t-test ($p < 0.05$).

Models	BIDMC+BMC (ID)						HK+I2CVB (OOD)						RUNMC+UCL (OOD)					
	Baseline			Baseline + **VarDeepPCA**			Baseline			Baseline + **VarDeepPCA**			Baseline			Baseline + **VarDeepPCA**		
	DSC	HD95	ASD	DSC	HD95	ASD	DSC	HD95	ASD	DSC	HD95	ASD	DSC	HD95	ASD	DSC	HD95	ASD
UNet	93.4	7.4	2.5	93.7	**4.6**	**1.9**	84.3	26.8	8.4	**88.0**	**7.3**	**2.9**	89.4	11.6	4.0	**90.6**	**6.7**	**2.6**
	3.9	13.8	2.4	3.4	**2.2**	**0.9**	7.4	23.9	6.2	**4.6**	**2.6**	**1.1**	5.9	13.6	4.1	**4.3**	**2.8**	**1.0**
AttnUNet	93.8	10.7	3.2	93.9	**4.7**	**1.9**	86.8	18.6	6.4	**89.4**	**6.7**	**2.7**	91.3	13.8	4.6	**91.9**	**5.8**	**2.4**
	2.7	21.1	4.8	2.5	**1.9**	**0.7**	9.9	22.6	8.5	**4.9**	**2.9**	**1.2**	4.6	22.2	6.3	**3.7**	**2.5**	**1.1**
ResUNet++	93.9	5.1	1.9	94.1	4.8	1.9	86.8	11.1	4.0	**87.6**	**7.5**	**3.1**	90.7	8.6	3.2	90.8	**6.7**	**2.7**
	2.6	3.5	1.0	2.4	2.0	0.8	7.0	11.1	3.1	**6.1**	**3.2**	**1.5**	4.5	8.5	2.3	4.3	**2.8**	**1.2**
DeepLabV3+	91.3	6.4	2.7	91.3	6.1	2.6	82.9	16.5	6.0	**86.0**	**8.2**	**3.5**	88.9	8.3	3.4	89.3	**6.7**	**3.0**
	3.2	2.6	1.3	3.2	2.4	1.0	13.3	17.6	6.2	**6.5**	**1.2**	**0.8**	4.6	6.4	1.8	4.0	**1.8**	**0.9**
BASNet	94.7	6.3	2.0	94.9	**4.2**	**1.6**	87.7	22.9	8.0	**91.7**	**5.5**	**2.3**	92.0	11.5	3.9	**92.7**	**5.3**	**2.2**
	2.5	12.0	2.5	2.1	**1.8**	**0.7**	9.6	27.5	10.0	**3.8**	**2.5**	**1.1**	5.1	19.5	6.1	**3.4**	**2.4**	**0.9**
SegAN	92.9	5.5	2.2	93.1	**5.1**	**2.0**	87.6	9.3	3.5	**88.1**	**7.4**	**3.0**	90.6	7.2	2.8	90.8	**6.2**	**2.5**
	3.2	2.7	1.0	3.1	**2.3**	**0.9**	4.7	7.1	2.0	**4.4**	**2.3**	**1.0**	3.9	6.1	1.4	3.8	**2.4**	**1.0**
MedSegDiff	90.3	24.5	7.4	**92.5**	**5.5**	**2.3**	79.1	47.5	16.8	**86.5**	**7.9**	**3.5**	72.4	73.2	28.1	**88.2**	**7.9**	**3.5**
	5.8	34.6	9.9	**3.6**	**2.4**	**1.2**	11.6	39.7	15.0	**4.5**	**2.1**	**1.1**	12.5	30.1	14.6	**3.8**	**2.2**	**1.1**
DSTransUNet	94.3	6.8	2.2	94.3	**4.6**	**1.9**	91.0	10.3	3.8	91.4	**6.2**	**2.5**	91.0	12.9	4.4	**91.5**	**6.1**	**2.6**
	2.5	10.5	2.0	2.4	**2.0**	**0.8**	4.9	14.2	4.9	4.1	**3.0**	**1.3**	5.1	17.8	4.9	**4.2**	**2.5**	**1.2**
VMUNet	91.5	5.9	2.6	91.8	5.6	2.4	84.9	8.7	4.1	**85.1**	**8.4**	3.9	87.8	7.6	3.5	88.0	7.4	3.5
	3.4	2.1	1.0	3.3	2.0	1.0	6.1	2.1	1.2	**5.2**	**1.9**	1.1	4.8	2.8	1.3	4.5	2.2	1.2
MedSAM	81.4	13.6	7.0	81.7	**7.9**	**3.7**	82.6	11.3	5.4	82.7	**7.0**	**3.5**	82.1	11.5	6.2	82.3	**8.1**	**3.6**
	5.7	4.4	2.5	5.8	**3.9**	**2.5**	5.5	3.5	2.2	5.6	**2.9**	**2.1**	5.5	4.1	2.4	5.6	**3.9**	**2.4**
PHISeg	93.1	5.2	2.1	93.3	5.1	2.1	87.6	7.5	3.2	87.8	7.5	3.3	90.7	6.4	2.8	90.9	6.3	2.8
	2.5	1.9	0.8	2.3	1.8	0.7	4.4	2.4	1.2	4.3	2.1	1.1	3.7	2.3	1.1	3.6	2.1	1.0
ProbUNet	91.4	11.3	3.6	**91.9**	**5.8**	**2.4**	83.3	18.7	6.5	**85.7**	**7.9**	**3.3**	86.7	10.9	4.2	**88.1**	**7.7**	**3.2**
	3.2	17.8	3.8	**2.8**	**2.0**	**0.9**	7.7	17.3	4.8	**4.9**	**2.3**	**1.1**	5.1	7.9	2.3	**4.4**	**2.3**	**1.0**
HierProbUNet	90.7	8.8	3.5	91.3	**6.7**	**2.7**	77.4	29.4	11.3	**84.8**	**7.9**	**3.7**	81.8	26.4	9.6	**88.0**	**7.9**	**3.3**
	3.8	7.6	2.3	3.5	**2.5**	**1.1**	10.7	13.2	5.6	**7.4**	**2.2**	**1.4**	6.9	12.8	4.8	**5.0**	**2.3**	**1.3**
SegCNN+DAE +TTA	92.3	5.3	2.3	92.7	5.2	2.1	87.1	8.9	3.4	**88.2**	**6.9**	**2.9**	89.4	7.3	2.9	89.5	6.8	2.8
	3.4	2.6	1.1	3.0	2.5	1.1	5.1	7.1	2.6	**4.7**	**2.6**	**1.6**	5.3	4.4	2.1	4.7	4.3	1.6
SegCNN+DAE +TTA+Atlas	93.7	4.7	1.9	93.7	4.7	1.9	88.5	8.0	3.1	**89.3**	**6.7**	**2.8**	90.5	6.8	2.6	90.5	6.6	2.6
	3.0	2.4	0.8	2.9	2.2	0.8	4.7	6.8	2.2	**4.2**	**2.3**	**1.0**	4.9	3.5	1.3	4.6	2.8	1.1
Mean of Baselines	92.0	8.5	3.1	92.3	**5.3**	**2.2**	86.1	14.5	5.4	**87.7**	**7.1**	**3.0**	88.5	12.7	4.9	**89.7**	**6.7**	**2.8**
	4.7	13.9	3.8	4.4	**2.5**	**1.2**	7.3	17.9	6.1	**5.4**	**2.6**	**1.4**	7.1	18.6	6.9	**5.1**	**2.9**	**1.4**

lines for the G1020 (OOD) as shown in Figure 7, and the ORIGA (OOD) dataset as shown in Figure 8; these were AttnUNet, MedSegDiff, and MedSAM (recall that MedSAM used oracle-prompting that is actually unfair to other methods). For variational models, PHISeg again exhibited superior robustness compared to ProbUNet and HierProbUNet. The figures also show that the per-pixel uncertainty maps produced by our VarDeepPCA method are much more aligned with the object anatomy as compared to those produced by the baselines.

Prostate Segmentation. Table 7 shows that the performance of a few large DNN models (SegAN, VMUNet, PHISeg) does not degrade much on OOD images with respect to the HD95 metric. Our VarDeepPCA plugin continues to deliver consistent improvements, and reduces the mean HD95 values for all baselines. In cases of severe degradation where baselines had high variance in HD95 values, e.g., MedSegDiff, our method also reduced the standard deviation, improving the stability of the results. We use a latent dimension of $K = 8$ for VarDeepPCA, as justified in our sensitivity analysis in Section 4.7. Indeed, for the OOD datasets, the mean HD95 values after employing the



(a) VMUNet (b) SegAN (c) SegCNN+DAE+TTA+Atlas
(d) PHISeg (e) ProbUNet (f) HierProbUNet
(g) Unc. for (d) PHISeg (h) Unc. for (e) ProbUNet (i) Unc. for (f) HierProbUNet
(j) Unc. for (d)+VarDeepPCA (k) Unc. for (e)+VarDeepPCA (l) Unc. for (f)+VarDeepPCA

Figure 10: **Results–Qualitative: Prostate Segmentation Restoration on RUNMC+UCL (OOD) data. (a)–(c)** Results on images for the best non-variational baselines. **(d)–(f)** Results on images for the variational baselines. **(g)–(i)** Uncertainty maps produced using variational baselines. **(j)–(l)** Uncertainty maps produced using VarDeepPCA when plugged into the associated baselines (d)–(f). Color scheme in (a)–(f): Baseline; Baseline+VarDeepPCA (Ours); Ground Truth. HD95 numbers in (a)–(f) indicate that the examples were representative of the test set, because the HD95 values were close to the mean HD95 reported in Table 7. NCC numbers in (g)–(l) indicate that the examples were representative of the test set, because the NCC values were close to the mean of the NCC reported in Table 9.

VarDeepPCA plugin reduce from an average (across all methods; before VarDeepPCA) within 12.7-14.5 to an average (across all methods; after VarDeepPCA) within 6.7-7.1, which is far more clinically acceptable as per our analysis in Section 4.3. We see minor improvements in the uncertainty calibration metrics from Table 9 for PHISeg, while there are major improvements for ProbUNet and HPUNet.

For qualitative analysis on OOD data, we selected the three best-performing non-variational baselines based on their mean HD95 metrics: (i) for the HK+I2CVB dataset, these were VMUNet, SegAN, and DSTransUNet (Figure 9); (ii) for the RUNMC+UCL dataset, these were VMUNet, SegAN, and DeepLabV3+ (Figure 10). Compared to myocardium segmentation and neuroretinal rim segmentation, VMUNet performed much better on the prostate dataset with relatively smaller HD95 values. However, VMUNet continued to exhibit high variance of the HD95 values. Our VarDeepPCA, when plugged into VMUNet, not only reduced the mean HD95 values but, more importantly, reduced the variance of the HD95 values significantly. The analysis of variational models revealed greater instability. While PHISeg's multi-scale architecture provided some robustness to OOD images, ProbUNet and HierProbUNet showed severe degradation. HierProbUNet, for instance, performed adequately on ID data but failed on both OOD datasets. As seen in Figure 9 and Figure 10, the uncertainty maps from these baselines often exhibited inconsistencies with the object's geometry, unlike the uncertainty maps produced by VarDeepPCA. Our method successfully addressed both issues, i.e., it improved the object segmentation maps and also improved the uncertainty maps.

Fetal Head Segmentation. Applying our VarDeepPCA framework yields consistent performance gains across both HC18 (ID) and FetalPlanes (OOD) datasets, with the improvements being most pronounced in the HD95 metric. This demonstrates that VarDeepPCA effectively corrects the boundary errors caused by noise and domain shift. For these experiments, we use a latent dimension of $K = 3$ for VarDeepPCA, as justified in our sensitivity analysis in Section 4.7. We also note that diffusion-based models like MedSegDiff, and models with boundary-aware loss function such as BASNet, outperformed the MedSAM foundational model (even when using oracle prompting) in this application. Table 8 shows that despite some baseline models (e.g., UNet and ResUNet++) achieving good (85%-90%) DSC scores (overlap based), their HD95 values (boundary based) are poor (around 40), even on the HC18 (ID) test set. This stems from the well-known limitation of DSC in segmenting large near-convex objects. Many baseline models exhibiting poor mean HD95 scores also showed high standard deviations (Table 8) in HD95, indicating inconsistent performance across images. In contrast, our method not only achieves a lower (better) mean HD95 but also simultaneously reduces the standard deviation, demonstrating higher accuracy and precision. In all cases, our method consistently restores degraded segmentation maps towards valid anatomical geometries. Indeed, for the OOD datasets, the mean HD95 values after employing the VarDeepPCA plugin reduce from an average (across all methods; before VarDeepPCA) within 11.5-12.7 to an average (across all methods; after VarDeepPCA) within 5.8-5.9, which is far more clinically acceptable as per our analysis in Section 4.3.

Table 8: **Results–Quantitative: Fetal Head Segmentation.** All models were trained on HC18 (ID) and evaluated on FetalPlanes (OOD). For each method-dataset combination, we report mean (top row) and standard deviation (bottom row, in gray) for DSC(↑), HD95(↓), and ASD(↓). Augmenting each baseline with our VarDeepPCA consistently improves performance. **Bold-font** values in the columns indicate a statistically significant improvement of the Baseline+VarDeepPCA method over the underlying Baseline method, using a one-tailed paired-sample t-test ($p < 0.05$).

Models	**HC18 (ID)**						**FetalPlanes (OOD)**					
	Baseline			Baseline + **VarDeepPCA**			Baseline			Baseline + **VarDeepPCA**		
	DSC	HD95	ASD	DSC	HD95	ASD	DSC	HD95	ASD	DSC	HD95	ASD
UNet	84.5	48.1	15.0	**96.9**	**5.2**	**2.2**	89.4	38.5	12.1	**96.5**	**5.2**	**2.4**
	12.9	10.9	3.9	**0.9**	**1.4**	**0.6**	8.6	18.6	5.8	**1.2**	**1.4**	**0.8**
AttnUNet	96.2	8.4	2.6	**97.0**	**5.2**	**2.1**	93.2	13.6	4.2	**96.6**	**5.1**	**2.3**
	6.6	13.0	3.4	**0.9**	**1.3**	**0.6**	8.8	16.0	3.9	**1.2**	**1.4**	**0.8**
ResUNet++	85.8	41.1	16.2	**95.8**	**6.2**	**3.0**	84.9	42.4	15.6	**96.0**	**5.7**	**2.8**
	7.4	15.8	7.1	**1.0**	**1.1**	**0.8**	7.3	14.1	5.7	**1.2**	**1.3**	**0.9**
DeepLabV3+	95.7	10.5	3.5	**96.8**	**5.4**	**2.3**	94.1	14.3	5.0	**96.5**	**5.5**	**2.5**
	3.0	6.7	2.3	**1.0**	**1.3**	**0.7**	2.4	6.6	2.2	**1.1**	**1.3**	**0.7**
BASNet	96.3	5.9	2.1	96.8	**5.3**	**1.9**	96.8	5.9	2.4	96.8	**5.0**	**2.3**
	1.5	8.1	2.3	1.0	**1.4**	**0.6**	1.2	4.0	1.1	1.1	**1.4**	**0.8**
SegAN	96.1	6.6	2.8	96.8	**5.2**	**2.2**	96.1	7.2	3.3	**96.5**	**5.2**	**2.4**
	2.3	4.7	1.8	1.0	**1.4**	**0.6**	1.5	3.6	1.1	**1.1**	**1.4**	**0.7**
MedSegDiff	96.2	5.3	2.7	97.0	**4.9**	**2.1**	96.0	7.2	2.7	**96.7**	**5.0**	**2.3**
	1.1	5.6	1.1	0.9	**1.4**	**0.6**	7.4	7.7	2.7	**1.0**	**1.4**	**0.7**
DSTransUNet	95.5	7.9	2.5	**96.4**	**5.6**	2.5	95.6	8.2	3.1	**96.3**	**5.4**	**2.5**
	1.4	3.7	1.0	**1.0**	**1.3**	0.6	1.7	3.6	1.1	**1.3**	**1.4**	**0.8**
VMUNet	96.0	6.4	2.4	**96.9**	**5.1**	**2.2**	95.6	8.7	3.4	**96.6**	**5.1**	**2.3**
	2.4	6.7	2.2	**0.9**	**1.4**	**0.6**	2.3	6.1	2.0	**1.1**	**1.4**	**0.7**
MedSAM	92.3	13.9	6.4	92.6	**10.8**	**5.7**	91.5	13.3	6.6	**91.8**	**10.5**	**6.1**
	4.5	7.8	4.0	4.3	**5.6**	**3.6**	4.6	7.2	3.9	**4.8**	**5.6**	**3.9**
PHISeg	96.0	5.3	2.1	**96.9**	**5.3**	**1.7**	95.6	7.2	2.9	**96.4**	**5.3**	**2.4**
	1.9	3.6	1.3	**0.9**	**1.3**	**0.6**	2.3	3.8	1.3	**1.1**	**1.4**	**0.7**
ProbUNet	95.7	18.6	5.0	95.9	**6.5**	**3.3**	93.9	14.1	4.9	**94.9**	**6.6**	**3.3**
	2.6	15.3	3.3	1.3	**1.2**	**0.8**	4.7	11.2	3.4	**1.1**	**0.9**	**0.7**
HierProbUNet	96.5	11.5	3.8	**96.8**	**5.3**	**2.2**	94.5	14.1	4.9	**96.4**	**5.3**	**2.4**
	2.0	7.5	2.3	**1.0**	**1.4**	**0.7**	2.8	7.7	2.2	**1.3**	**1.5**	**0.8**
SegCNN+DAE +TTA	96.1	8.9	2.7	96.2	**5.3**	**2.3**	95.3	8.5	3.1	**96.1**	**5.7**	**2.6**
	3.1	10.6	3.3	1.5	**1.8**	**0.9**	3.1	6.9	2.3	**1.8**	**1.6**	**1.1**
SegCNN+DAE +TTA+Atlas	96.9	8.6	2.5	96.9	**5.1**	**2.2**	95.8	8.0	2.8	**96.4**	**5.2**	**2.4**
	2.9	10.3	2.9	1.1	**1.4**	**0.7**	2.8	6.2	1.7	**1.3**	**1.4**	**0.8**
Mean of Baselines	94.8	11.5	4.1	**96.4**	**5.9**	**2.5**	94.2	12.7	4.7	**96.0**	**5.8**	**2.8**
	5.7	13.6	4.4	**2.2**	**2.9**	**1.8**	5.5	12.9	4.3	**2.4**	**2.8**	**1.9**

From Table 10, we see that the models ProbUNet and HPUNet are not well calibrated and do not provide good uncertainty estimates, whereas, PHISeg is relatively well calibrated and provides better uncertainty estimates.

For qualitative analysis on the FetalPlanes (OOD) dataset, we show examples from three best-performing non-variational baselines: BASNet, SegAN, and MedSegDiff. These baselines frequently produce severe mispredictions, with segmentation boundaries extending well beyond the anatomical boundaries (Figure 11). In contrast, our method successfully restores these degraded masks to anatomically valid geometries, much closer to the ground truth. Regarding the variational models, PHISeg again performs better than ProbUNet and HierProbUNet on the OOD dataset, probably because PHISeg leverages a more sophisticated larger backbone DNN than UNet. Figure 11 also shows that our method generates superior uncertainty maps; the uncertainty is suitably localized and respects the geometry of the fetal head, unlike the more diffuse or inaccurate uncertainty maps from the other variational baselines.

Table 9: **Results–Quantitative: Prostate – Measuring Calibration between Per-Pixel Segmentation Uncertainty and Per-Pixel Segmentation Error.** All models were trained on BIDMC+BMC (ID), and evaluated on HK+I2CVB and RUNMC+UCL (both OOD). For each method-dataset combination, we report the mean (top row) and standard deviation (bottom row; in gray) for NCC (↑), US (↑), and TACE (↓) metrics. Augmenting each baseline with our VarDeepPCA shows better calibration. **Bold-font** values in the columns indicate a statistically significant improvement of the Baseline+VarDeepPCA method over the underlying Baseline method, using a one-tailed paired-sample t-test ($p < 0.05$).

Models	BIDMC+BMC (ID)						HK+I2CVB (OOD)						RUNMC+UCL (OOD)					
	Baseline			Baseline + **VarDeepPCA**			Baseline			Baseline + **VarDeepPCA**			Baseline			Baseline + **VarDeepPCA**		
	NCC	US	TACE	NCC	US	TACE	NCC	US	TACE	NCC	US	TACE	NCC	US	TACE	NCC	US	TACE
PHISeg	0.66	0.76	0.39	**0.69**	**0.82**	**0.12**	0.57	0.63	0.47	**0.63**	**0.69**	**0.14**	0.45	0.66	0.53	**0.52**	**0.73**	**0.09**
	0.14	0.04	0.26	**0.10**	**0.04**	**0.24**	0.11	0.03	0.23	**0.08**	**0.04**	**0.24**	0.05	0.04	0.17	**0.03**	**0.04**	**0.12**
ProbUNet	0.45	0.69	0.44	**0.54**	**0.73**	**0.14**	0.40	0.59	0.50	**0.45**	**0.65**	**0.12**	0.51	0.68	0.46	**0.55**	**0.74**	**0.08**
	0.08	0.03	0.31	**0.08**	**0.04**	**0.25**	0.06	0.03	0.32	**0.06**	**0.03**	**0.23**	0.03	0.03	0.25	**0.02**	**0.03**	**0.12**
HPUNet	0.51	0.69	0.43	**0.61**	**0.79**	**0.16**	0.54	0.64	0.65	**0.59**	**0.75**	**0.19**	0.47	0.65	0.46	**0.56**	**0.74**	**0.12**
	0.09	0.03	0.29	**0.08**	**0.03**	**0.24**	0.07	0.03	0.25	**0.06**	**0.02**	**0.18**	0.04	0.04	0.22	**0.03**	**0.03**	**0.15**

4.7 Sensitivity Analysis

We evaluated the sensitivity of our VarDeepPCA plugin to the choice of the latent dimension K and the training set size S. We selected VMUNet and PHISeg as the baselines specifically because our earlier experiments demonstrated their relative robustness to OOD data in all of the applications. We also selected UNet as an early DNN method. We report our analysis on a pooled test set comprising the ID test set and one representative OOD dataset for each application.

Changing Latent Dimension K Within VarDeepPCA. We trained different versions of VarDeepPCA (training set size $S = 150$) across a range of K values for each application. For myocardium segmentation on CAP (ID) and ACDC (OOD) datasets, the baselines yielded a median HD95 of approximately 10. Our VarDeepPCA models with $K \in \{12, 16, 20\}$ consistently reduced this to a mean of 5–6, as shown in Figure 12(a). We find the performance to be relatively insensitive to to small changes in K. We observed similar trends for all other datasets, i.e., for the neuroretinal rim segmentation on MAGRABI (ID) + ORIGA (OOD) datasets (Figure 12(b)), for prostate segmentation on BIDMC+BMC (ID) + HK+I2CVB (OOD) datasets (Figure 12(c)) for fetal head segmentation on HC18 (ID) + FetalPlanes (OOD) datasets (Figure 12(d)). Across all applications and all tested latent dimensions, our VarDeepPCA plugin consistently improved the baseline segmentations, demonstrating significant robustness regardless of the specific K value chosen.

Changing Size of Training Set S. To analyze sensitivity to training-data availability, we used training sets

Table 10: **Results–Quantitative: Fetal Head – Measuring Calibration between Per-Pixel Segmentation Uncertainty and Per-Pixel Segmentation Error.** All models were trained on HC18 (ID), and evaluated on FetalPlanes (OOD). For each method-dataset combination, we report the mean (top row) and standard deviation (bottom row; in gray) for NCC (↑), US (↑), and TACE (↓) metrics. Augmenting each baseline with our VarDeepPCA shows better calibration. **Bold-font** values in the columns indicate a statistically significant improvement of the Baseline+VarDeepPCA method over the underlying Baseline method, using a one-tailed paired-sample t-test ($p < 0.05$).

Models	HC18 (ID)						FetalPlanes (OOD)					
	Baseline			Baseline + **VarDeepPCA**			Baseline			Baseline + **VarDeepPCA**		
	NCC	US	TACE	NCC	US	TACE	NCC	US	TACE	NCC	US	TACE
PHISeg	0.67	0.69	0.34	**0.73**	**0.75**	**0.14**	0.49	0.71	0.45	**0.55**	**0.78**	**0.08**
	0.06	0.03	0.04	**0.03**	**0.03**	**0.03**	0.05	0.02	0.05	**0.04**	**0.00**	**0.04**
ProbUNet	0.54	0.72	0.37	**0.68**	**0.82**	**0.15**	0.34	0.65	0.44	**0.45**	**0.75**	**0.11**
	0.05	0.03	0.08	**0.03**	**0.04**	**0.04**	0.07	0.02	0.03	**0.05**	**0.01**	**0.03**
HPUNet	0.46	0.71	0.41	**0.54**	**0.77**	**0.25**	0.35	0.65	0.46	**0.53**	**0.74**	**0.14**
	0.06	0.02	0.04	**0.03**	**0.03**	**0.03**	0.06	0.01	0.04	**0.05**	**0.00**	**0.03**

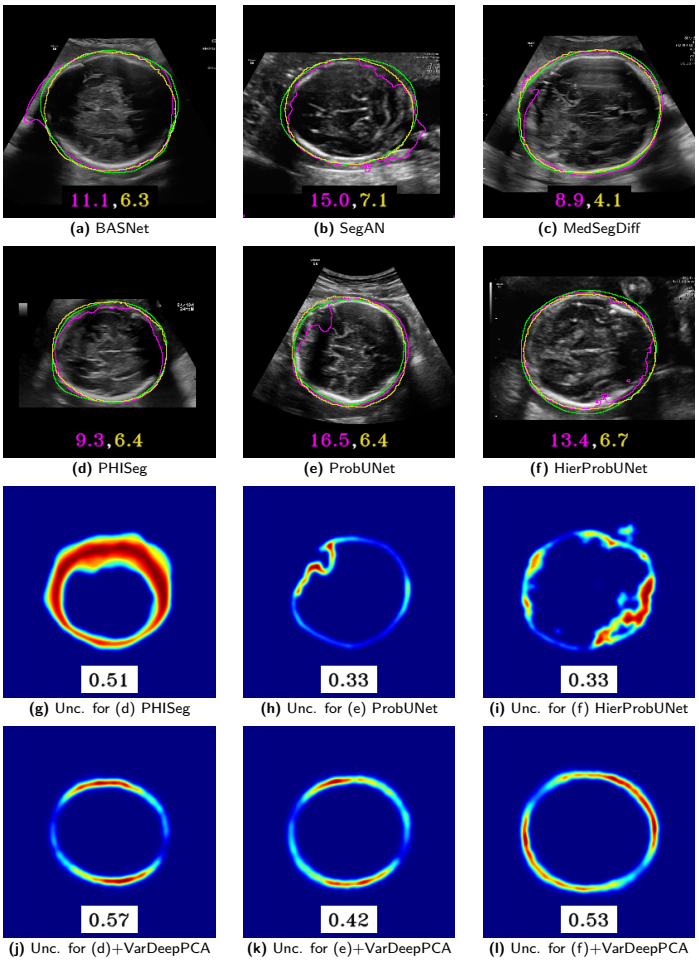


Figure 11: **Results–Qualitative: Fetal Head Segmentation Restoration on FetalPlanes (OOD) data. (a)–(c)** Results on images for the best non-variational baselines. **(d)–(f)** Results on images for the variational baselines. **(g)–(i)** Uncertainty maps produced using variational baselines. **(j)–(l)** Uncertainty maps produced using VarDeepPCA when plugged into the associated baselines (d)–(f). Color scheme in (a)–(f): Baseline; Baseline+VarDeepPCA (Ours); Ground Truth. HD95 numbers in (a)–(f) indicate that the examples were representative of the test set, because the HD95 values were close to the mean HD95 reported in Table 8. NCC numbers in (g)–(l) indicate that the examples were representative of the test set, because the NCC values were close to the mean of the NCC reported in Table 10.

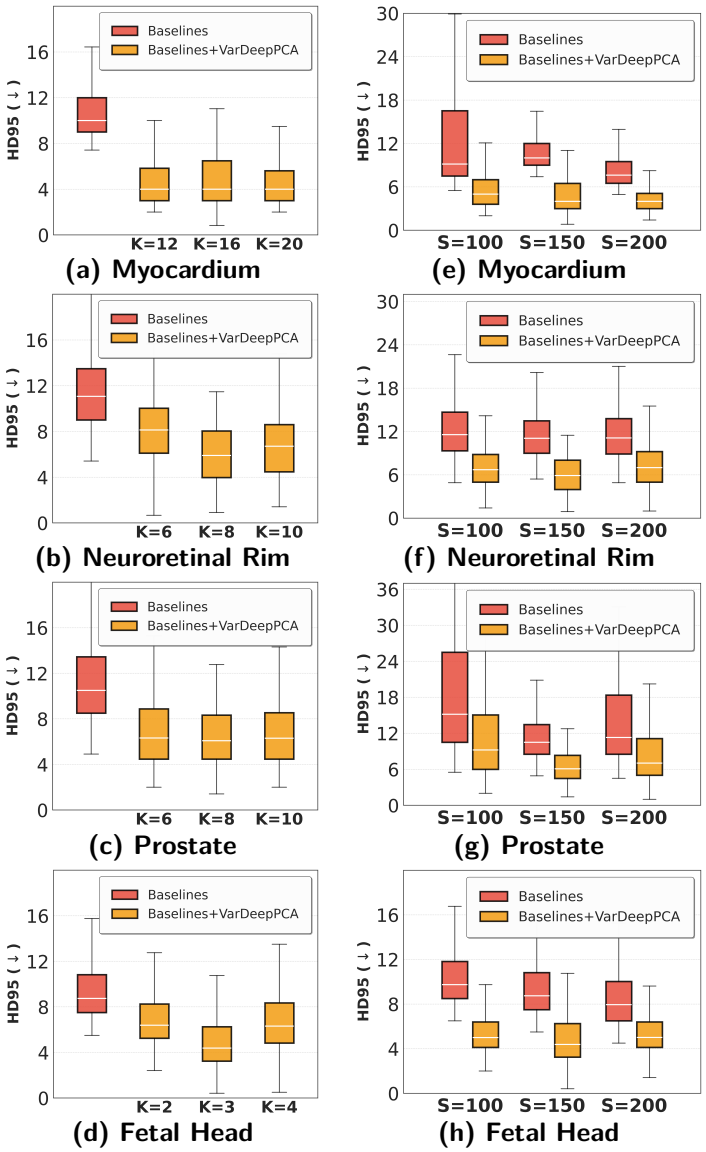


Figure 12: **Sensitivity of Results to Choice of Latent Dimension K and Training Set Size S.** We pooled results using three baselines (UNet, VMUNet, PHISeg) and their corresponding versions with VarDeepPCA plugged in. **(a)–(d)** Boxplots of HD95 values for varying latent dimension K. **(e)–(h)** Boxplots of HD95 values for varying training-set sample size S. HD95 values on pooled ID and OOD test sets for: **(a,e)** Myocardium (CAP + ACDC); **(b,f)** Neuroretinal Rim (MAGRABI + G1020); **(c,g)** Prostate (BIDMC+BMC + HK+I2CVB); **(d,h)** Fetal Head (HC18 + FetalPlanes).

of three sizes: S_{100}, S_{150}, and S_{200}. We designed them to be nested such that $S_{100} \subset S_{150} \subset S_{200}$. The baselines (UNet, VMUNet, PHISeg) were trained on each subset and evaluated on the pooled ID and OOD test set (Figure 12(e)–(h)). Across all applications and datasets, we find the performances of the methods to be quite insensitive to small changes in S. Moreover, across all training-set sample sizes, our plugin consistently and significantly improved over the baselines.

5. Conclusion

We have presented VarDeepPCA, a novel framework designed to restore degraded segmentation masks produced by existing DNNs on OOD images. By explicitly learning the principal modes of anatomical geometric variation from a small ID dataset of segmentation maps, our framework neither requires access to any OOD data nor to any medi-

cal/acquired image. Our novel variational learning framework leverages a reinterpretation of the softmax mapping to implicitly perform exact distribution modeling, thereby enabling computationally efficient, sampling-free learning and inference. Our framework is characterized by a lightweight architecture (ranging from 1.02M to 2.72M parameters depending on K), modality independence, and the ability to generate reliable/geometrically-consistent uncertainty estimates. Extensive empirical validation across 14 publicly available datasets and 15 DNN segmenters confirms that when VarDeepPCA is plugged in to existing DNN methods, it consistently improves existing methods in OOD object segmentation as well as uncertainty estimation. Indeed, the VarDeepPCA plugin may also be extended to improve poor OOD-image segmentations resulting from traditional non-DNN methods. While some of the baselines exhibit robustness sporadically on specific datasets, our framework provides a much more reliable and consistent improvement in terms of boundary-delineation accuracy and the plausibility of anatomical geometry.

We acknowledge a few constraints of our framework. First, the method relies on consistency of geometry of the anatomical objects of interest; therefore, it is inherently unsuitable for segmenting pathologies with highly non-uniform or amorphous/unpredictable geometries, e.g., lesions and tumors. Second, if a segmentation map (input to VarDeepPCA) lies within the learned distribution of valid geometries but is still incorrect, then VarDeepPCA would be unable to improve the segmentation. Finally, in scenarios where the error in the segmentation produced by an existing method is extremely large, then our (or any such other) framework may be unable to restore the correct segmentation. Nevertheless, even in such cases, VarDeepPCA succeeds in projecting the segmentation map to one with a valid anatomical geometry. Future work may include extensions of VarDeepPCA to multi-class segmentation problems in 2D and 3D images. Future work may also explore the impact of patient demographics and clinical metadata on model generalization and uncertainty.

Acknowledgments

Supported by the Prime Minister's Research Fellowship from the Government of India.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

We declare we don't have conflicts of interest.

Data availability

To facilitate reproducibility, the source code, datasets, and pre-trained model weights will be made publicly available upon publication.

References

S. Adiga, J. Dolz, and H. Lombaert. Anatomically-aware uncertainty for semi-supervised image segmentation. *Med. Image Anal.*, 91:103011, 2024.

A. Almazroa, R. Burman, K. Raahemifar, and V. Lakshminarayanan. Optic disc and optic cup segmentation methodologies for glaucoma image detection: a survey. *J. Ophthalmol.*, 2015(1):1–26, 2015.

A. Almazroa, S. Alodhayb, E. Osman, E. Ramadan, M. Hummadi, M. Dlaim, M. Alkatee, K. Raahemifar, and V. Lakshminarayanan. Retinal fundus images for glaucoma analysis: the RIGA dataset. In *Medical Imaging: Imaging Informatics for Healthcare, Research, and Applications*, volume 10579, pages 55–62, 2018.

O. A. M. F. Alnaggar, B. N. Jagadale, M. A. N. Saif, O. A. M. Ghaleb, A. A. Q. Ahmed, H. A. A. Aqlan, and H. D. E. Al-Ariki. Efficient artificial intelligence approaches for medical image processing in healthcare: comprehensive review, taxonomy, and analysis. *Artif. Intell. Rev.*, 57(8):221, 2024.

P. Andreini, S. Bonechi, M. Bianchini, A. Mecocci, and F. Scarselli. Image generation by GAN and style transfer for agar plate image segmentation. *Comput. Methods Programs Biomed.*, 184:105268, 2020.

A. Andreopoulos and J. Tsotsos. Efficient and generalizable statistical models of shape and appearance for analysis of cardiac MRI. *Med. Image Anal.*, 12(3):335–57, 2008.

S Awate, S Garg, and R Jena. Estimating uncertainty in MRF-based image segmentation: A perfect-MCMC approach. *Med. Imag. Analysis*, 55:181–96, 2019.

M. Bajwa, G. Singh, W. Neumeier, M. Malik, A. Dengel, and S. Ahmed. G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection. In *Int. Joint Conf. Neural Networks*, pages 1–7, 2020.

J. O. Barentsz, J. Richenberg, R. Clements, P. Choyke, S. Verma, G. Villeirs, O. Rouviere, V. Logager, and

J. J. Fütterer. ESUR prostate MR guidelines 2012. *Eur. Radiol.*, 22:746–57, 2012.

M. Bateson, H. Kervadec, J. Dolz, H. Lombaert, and I. B. Ayed. Source-free domain adaptation for image segmentation. *Med. Image Anal.*, 82:102617, 2022.

C. F. Baumgartner, K. C. Tezcan, K. Chaitanya, A. M. Hötker, U. J. Muehlematter, K. Schawkat, A. S. Becker, O. Donati, and E. Konukoglu. PHISeg: Capturing uncertainty in medical image segmentation. In *Med. Image Comput. Comput.-Assist. Interv.*, pages 119–127. Springer, 2019.

O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. Gonzalez Ballester, G. Sanroma, S. Napel, S. Petersen, G. Tziritas, E. Grinias, M. Khened, V. A. Kollerathu, G. Krishnamurthi, M.-M. Rohé, X. Pennec, M. Sermesant, F. Isensee, P. Jäger, K. H. Maier-Hein, P. M. Full, I. Wolf, S. Engelhardt, C. F. Baumgartner, L. M. Koch, J. M. Wolterink, I. Išgum, Y. Jang, Y. Hong, J. Patravali, S. Jain, O. Humbert, and P.-M. Jodoin. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Trans. Med. Imag.*, 37(11):2514–25, 2018.

N. Bloch, A. Madabhushi, H. Huisman, J. Freymann, J. Kirby, M. Grauer, A. Enquobahrie, C. Jaffe, L. Clarke, and K. Farahani. Nci-isbi 2013 challenge: automated segmentation of prostate structures. *The Cancer Imaging Archive*, 370(6):5, 2015.

X. P. Burgos-Artizzu, D. Coronado-Gutiérrez, B. Valenzuela-Alcaraz, E. Bonet-Carne, E. Eixarch, F. Crispi, and E. Gratacós. Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. *Scientific Reports*, 10(1): 10200, 2020.

A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2): 125, 2020.

J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang, M. P. Lungren, S. Zhang, L. Xing, L. Lu, A. Yuille, and Y. Zhou. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Med. Image Anal.*, 97:103280, 2024.

L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Eur. Conf. Comput. Vis.*, volume 11211, pages 833–50, 2018a.

L. C. Chen, G. Papandreou, I. Kokkinos, K. P. Murphy, and A. L. Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40:834–48, 2018b.

K. Cho. Simple sparsification improves sparse denoising autoencoders in denoising highly corrupted images. In *Int. Conf. Mach. Learn.*, pages 432–40. PMLR, 2013.

F. G. Claus, H. Hricak, and R. R. Hattery. Pretreatment evaluation of prostate cancer: role of MR imaging and 1H MR spectroscopy. *Radiographics*, 24:S167–80, 2004.

T. M. Cover. *Elements of information theory*. John Wiley & Sons, 1999.

A. Creswell and A. Bharath. Denoising adversarial autoencoders. *IEEE Transactions on Neural Networks and Learning Systems*, 30(4):968–84, 2018.

J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 248–55, 2009.

L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Int. Conf. Learn. Represent.*, pages 1–21, 2021.

F. Epstein. MRI of left ventricular function. *J. Nucl. Cardiol.*, 14(5):729–44, 2007.

Y. Fang, P. T. Yap, W. Lin, H. Zhu, and M. Liu. Source-free unsupervised domain adaptation: A survey. *Neural Networks*, 174:106230, 2024.

S. Farquhar and Y. Gal. What 'Out-of-distribution' is and is not. In *Adv. Neural Inform. Process. Syst. ML Safety Workshop*, pages 1–7, 2022.

P. Fischer, K. Thomas, and C. F. Baumgartner. Uncertainty estimation and propagation in accelerated MRI reconstruction. In *Int. Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 84–94. Springer, 2023.

E. Fuchs and A. Duane. *Text-book of Ophthalmology*. JB Lippincott, 1908.

A. V. Gaikwad and S. Awate. Deep monte-carlo EM for semantic segmentation using weakly-and-semi-supervised learning using very few expert segmentations. *Journal of Machine Learning for Biomedical Imaging*, 2:717–60, 2024.

A. V. Gaikwad and S. P. Awate. Deep MCEM for weakly-supervised learning to jointly segment and recognize objects using very few expert segmentations. In *Inf. Process. Med. Imaging*, pages 624–36. Springer, 2021.

A. V. Gaikwad, H. Varma, and S. P. Awate. Deep variational segmentation of topology-constrained object sets, with correlated uncertainty models, for robustness to degradations. In *IEEE Int. Conf. Image Process.*, pages 2195–99. IEEE, 2023.

A. Galdran, G. Carneiro, and M. A. González Ballester. On the optimal combination of cross-entropy and soft dice losses for lesion segmentation with out-of-distribution robustness. In *Diabetic Foot Ulcers Grand Challenge*, pages 40–51. Springer, 2022.

J. Gao, Q. Lao, P. Liu, H. Yi, Q. Kang, Z. Jiang, X. Wu, K. Li, Y. Chen, and L. Zhang. Anatomically guided cross-domain repair and screening for ultrasound fetal biometry. *IEEE J. Biomed. Health Inform.*, 27(10):4914–25, 2023a.

S. Gao, H. Zhou, Y. Gao, and X. Zhuang. Bayeseg: Bayesian modeling for medical image segmentation with interpretable generalizability. *Med. Image Anal.*, 89:102889, 2023b.

A. Giri, G. P. T. Choi, and L. Kumar. Open and closed anatomical surface description via hemispherical area-preserving map. *Signal Process.*, 180:107867, 2021.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Adv. Neural Inform. Process. Syst.*, pages 2672–80, 2014.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *Int. Conf. Mach. Learn.*, pages 1321–30, 2017.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

J. Healy and L. McInnes. Uniform manifold approximation and projection. *Nature Reviews Methods Primers*, 4(1): 82, 2024.

D. Hendrycks, K. Lee, and M. Mazeika. Using pre-training can improve model robustness and uncertainty. In *Int. Conf. Mach. Learn.*, pages 2712–2721. PMLR, 2019.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Adv. Neural Inform. Process. Syst.*, volume 30, 2017.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Adv. Neural Inform. Process. Syst.*, volume 33, pages 6840–51, 2020.

J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7132–41, 2018.

Y. Huang, X. Yang, L. Liu, H. Zhou, A. Chang, X. Zhou, R. Chen, J. Yu, J. Chen, C. Chen, S. Liu, H. Chi, X. Hu, K. Yue, L. Li, V. Grau, D.-P. Fan, F. Dong, and D. Ni. Segment anything model for medical images? *Med. Image Anal.*, 92:103061, 2024.

D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge. Comparing images using the hausdorff distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(9):850–63, 1993.

A. Jacob, P. Sharma, and D. Rueckert. Deep conditional shape models for 3D cardiac image segmentation. In *Statistical Atlases and Computational Models of the Heart. Regular and CMRxRecon Challenge Papers*, volume 14507 of *Lecture Notes Comput. Sci.*, pages 44–54. Springer, 2023.

A. Jacob, P. Sharma, and D. Rueckert. DCSM 2.0: Deep conditional shape models for data efficient segmentation. In *IEEE Int. Symp. Biomed. Imaging*, pages 1–4, 2024.

S. Jadon. A survey of loss functions for semantic segmentation. In *IEEE Conf. on Computational Intell. in Bioinformatics and Computational Biology*, pages 1–7, 2020.

R. Jena and S. P. Awate. A Bayesian neural net to segment images with uncertainty estimates and good calibration. In *Int. Conf. on Inf. Process. in Medical Imaging*, pages 3–15, 2019.

D. Jha, P. Smedsrud, M. Riegler, D. Johansen, T. Lange, P. Halvorsen, and H. Johansen. ResUNet++: An advanced architecture for medical image segmentation. *IEEE Int. Symposium on Multimedia*, pages 225–30, 2019.

R. El Jurdi, C. Petitjean, P. Honeine, V. Cheplygina, and F. Abdallah. High-level prior-based loss functions for medical image segmentation: A survey. *Comput. Vis. Image Underst.*, 210:103248, 2021.

A. Kadish, D. Bello, J. Finn, R. Bonow, A. Schaechter, H. Subacius, C. Albert, J. Daubert, C. Fonseca, and

J. Goldberger. Rationale and design for the defibrillators to reduce risk by magnetic resonance imaging evaluation DETERMINE trial. *Journal of Cardiovascular Electrophysiology*, 20(9):982–87, 2009.

R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

N. Karani, E. Erdil, K. Chaitanya, and E. Konukoglu. Test-time adaptable neural networks for robust medical image segmentation. *Med. Image Anal.*, 68:101907, 2021.

A. Kazerouni, E. Khodapanah Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof. Diffusion models for medical image analysis: A comprehensive survey. *Med. Image Anal.*, 88:102846, 2023.

H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. Ben Ayed. Boundary loss for highly unbalanced segmentation. *Med. Image Anal.*, 67:101851, 2021.

D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Int. Conf. Learn. Represent.*, pages 1–15, 2015.

D. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *Int. Conf. Learn. Represent.*, pages 1–14, 2014.

A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick. Segment anything. In *Int. Conf. Comput. Vis.*, pages 4015–26, 2023.

S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. M. Eslami, D. Jimenez Rezende, and O. Ronneberger. A probabilistic U-Net for segmentation of ambiguous images. *Adv. Neural Inform. Process. Syst.*, 31, 2018.

S. A. A. Kohl, B. Romera-Paredes, K. H. Maier-Hein, D. Jimenez Rezende, S. M. Eslami, P. Kohli, A. Zisserman, and O. Ronneberger. A hierarchical probabilistic U-Net for modeling multi-scale ambiguities. *arXiv preprint arXiv:1905.13077*, 2019.

F. Laakom, J. Raitoharju, A. Iosifidis, and M. Gabbouj. Reducing redundancy in the bottleneck representation of autoencoders. *Pattern Recognit. Lett.*, 178:202–8, 2024.

F. Langkilde, P. Masaba, L. Edenbrandt, M. Gren, A. Halil, M. Hellström, M. Larsson, A. A. Naeem, J. Wallström, S. E. Maier, and F. Jäderling. Manual prostate MRI segmentation by readers with different experience: a study of the learning progress. *Eur. Radiol.*, 34(7):4801–9, 2024.

A. J. Larrazabal, C. Martínez, B. Glocker, and E. Ferrante. Post-DAE: anatomically plausible segmentation via post-processing with denoising autoencoders. *IEEE Trans. Med. Imag.*, 39(12):3813–3820, 2020.

G. Lemaître, R. Martí, J. Freixenet, J. C. Vilanova, P. M. Walker, and F. Meriaudeau. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: a review. *Comput. Biol. Med.*, 60: 8–31, 2015.

J. Lennartz and T. Schultz. Segmentation distortion: Quantifying segmentation uncertainty under domain shift via the effects of anomalous activations. In *Med. Image Comput. Comput.-Assist. Interv.*, pages 316–25. Springer, 2023.

B. Li, Y. Liu, C. Occleshaw, B. Cowan, and A. Young. In-line automated tracking for ventricular function with magnetic resonance imaging. *JACC Cardiovascular Imaging*, 3(8): 860–66, 2010.

X. Li, H. Ding, H. Yuan, W. Zhang, J. Pang, G. Cheng, K. Chen, Z. Liu, and C. Loy. Transformer-based visual segmentation: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46:10138–63, 2024.

J. Liang, R. He, and T. Tan. A comprehensive survey on test-time adaptation under distribution shifts. *Int. J. Comput. Vis.*, 133(1):31–64, 2025.

A. Lin, B. Chen, J. Xu, Z. Zhang, G. Lu, and D. Zhang. DS-TransUNet: Dual swin transformer U-Net for medical image segmentation. *IEEE Trans. Instr. Meas.*, 71:1–15, 2021.

G. Litjens, R. Toth, W. van de Ven, C. Hoeks, S. Kerkstra, B. van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang, R. Strand, F. Malmberg, Y. Ou, C. Davatzikos, M. Kirschner, F. Jung, J. Yuan, W. Qiu, Q. Gao, P. Edwards, B. Maan, F. van der Heijden, S. Ghose, J. Mitra, J. Dowling, D. Barratt, H. Huisman, and A. Madabhushi. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. *Med. Image Anal.*, 18(2): 359–73, 2014.

J. Liu, H. Yang, H.-Y. Zhou, Y. Xi, L. Yu, Y. Yu, Y. Liang, G. Shi, S. Zhang, H. Zheng, and S. Wang. Swin-UMamba: Mamba-based unet with ImageNet-based pretraining. In *Med. Image Comput. Comput.-Assist. Interv.*, pages 615–25. Springer, 2024.

Q. Liu, C. Chen, J. Qin, Q. Dou, and P. Heng. FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1013–23, 2021.

S. Lu. Accurate and efficient optic disc detection and segmentation by a circular transformation. *IEEE Trans. Med. Imag.*, 30(12):2126–33, 2011.

J. Ma, F. Li, and B. Wang. U-Mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

A. Mehrtash, W. M. Wells, C. M. Tempany, P. Abolmaesumi, and T. Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Trans. Med. Imag.*, 39(12):3868–78, 2020.

R. Mehta, A. Filos, U. Baid, C. Sako, R. McKinley, M. Rebsamen, K. Dätwyler, R. Meier, P. Radojewski, G. K. Murugesan, S. Nalawade, C. Ganesh, B. Wagner, F. F. Yu, B. Fei, A. J. Madhuranthakam, J. A. Maldjian, L. Daza, C. Gómez, P. Arbeláez, C. Dai, S. Wang, H. Reynaud, Y. Mo, E. Angelini, Y. Guo, W. Bai, S. Banerjee, L. Pei, M. Ak, S. Rosas-González, I. Zemmoura, C. Tauber, M. H. Vu, T. Nyholm, T. Löfstedt, L. M. Ballestar, V. Vilaplana, H. McHugh, G. M. Talou, A. Wang, J. Patel, K. Chang, K. Hoebel, M. Gidwani, N. Arun, S. Gupta, M. Aggarwal, P. Singh, E. R. Gerstner, J. Kalpathy-Cramer, N. Boutry, A. Huard, L. Vidyaratne, M. M. Rahman, K. M. Iftekharuddin, J. Chazalon, E. Puybareau, G. Tochon, J. Ma, M. Cabezas, X. Llado, A. Oliver, L. Valencia, S. Valverde, M. Amian, M. Soltaninejad, A. Myronenko, A. Hatamizadeh, X. Feng, Q. Dou, N. Tustison, C. Meyer, N. A. Shah, S. Talbar, M.-A. Weber, A. Mahajan, A. Jakab, R. Wiest, H. M. Fathallah-Shaykh, A. Nazeri, M. Milchenko, D. Marcus, A. Kotrotsou, R. Colen, J. Freymann, J. Kirby, C. Davatzikos, B. Menze, S. Bakas, Y. Gal, and T. Arbel. QU-BraTS: MICCAI BraTS 2020 challenge on quantifying uncertainty in brain tumor segmentation–analysis of ranking scores and benchmarking results. *Journal of Machine Learning for Biomedical Imaging*, 2022:1–60, 2022.

S. Mohamed, M. Rosca, M. Figurnov, and A. Mnih. Monte carlo gradient estimation in machine learning. *Journal of Machine Learning Research*, 21(132):1–62, 2020.

V. Nagabotu and A. Namburu. Precise segmentation of fetal head in ultrasound images using improved U-Net model. *ETRI Journal*, 46(3):526–37, 2024.

M. Nawaz, A. Uvaliyev, K. Bibi, H. Wei, S. M. D. Abaxi, A. Masood, P. Shi, H. Ho, and W. Yuan. Unraveling the complexity of optical coherence tomography image segmentation using machine and deep learning techniques: A review. *Computerized Medical Imaging and Graphics*, 108:102269, 2023.

J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran. Measuring calibration in deep learning. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, volume 2(7), 2019.

O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, S. Cook, A. Marvao, T. Dawes, D. O'Regan, B. Kainz, B. Glocker, and D. Rueckert. Anatomically Constrained Neural Networks (ACNNs): Application to cardiac image enhancement and segmentation. *IEEE Trans. Med. Imag.*, 37(2):384–95, 2017.

O. Oktay, J. Schlemper, L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Hammerla, B. Kainz, B. Glocker, and D. Rueckert. Attention U-Net: Learning where to look for the pancreas. In *Conf. Med. Imaging Deep Learn.*, pages 1–10, 2018.

C. Ouyang, C. Chen, S. Li, Z. Li, C. Qin, W. Bai, and D. Rueckert. Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Trans. Med. Imag.*, 42(4):1095–1106, 2022.

N. Painchaud, Y. Skandarani, T. Judge, O. Bernard, A. Lalande, and P. Jodoin. Cardiac segmentation with strong anatomical guarantees. *IEEE Trans. Med. Imag.*, 39(11): 3703–13, 2020.

J. Pal. Holistic network for quantifying uncertainties in medical images. In *International MICCAI brainlesion workshop*, pages 560–69. Springer, 2021.

J. Pal and S. Awate. Convex segments for convex objects using DNN boundary tracing and graduated optimization. In *Med. Image Comput. Comput.-Assist. Interv.*, pages 91–101, 2024a.

J. Pal and D. Mj. Improving multi-scale attention networks: Bayesian optimization for segmenting medical images. *The Imaging Science Journal*, 71:33–49, 2023.

J. Pal, A. Bhattacharyea, D. Banerjee, and B. T. Maharaj. Advancing instance segmentation and wbc classification in peripheral blood smear through domain adaptation: A study on pbc and the novel rv-pbs datasets. *Expert Syst. Appl.*, 249:123660, 2024.

J. B. Pal and S. P. Awate. A hard convex-shape constraint in DNNs for object segmentation. In *IEEE Int. Conf. Image Process.*, pages 2074–80. IEEE, 2024b.

J. B. Pal, S. Welling, H. Saini, and S. P. Awate. Reviving poor object segmentations in OOD medical images using Variational-Deep-PCA modeling on segmentation maps with sampling-free learning. In *IEEE Winter Conf. Appl. Comput. Vis.*, pages 9346–55, 2025.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga,

A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An imperative style, high-performance deep learning library. *Adv. Neural Inform. Process. Syst.*, 32, 2019.

P. Peng, K. Lekadir, A. Gooya, L. Shao, S. E. Petersen, and A. F. Frangi. A review of heart chamber segmentation for structural and functional analysis using cardiac magnetic resonance imaging. *Magn. Reson. Mater. Phy.*, 29:155–95, 2016.

C. Petitjean and J. N. Dacher. A review of segmentation methods in short axis cardiac MR images. *Med. Image Anal.*, 15(2):169–84, 2011.

F. Pinto, P. Torr, and P. K. Dokania. Are vision transformers always more robust than convolutional neural networks? In *NeurIPS 2021 Workshop on distribution shifts: connecting methods and applications*, 2021.

X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jägersand. BASNet: Boundary-aware salient object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7479–89, 2019.

A. Rahman, J. M. J. Valanarasu, I. Hacihaliloglu, and V. M. Patel. Ambiguous medical image segmentation using diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11536–46, 2023.

A. Razavi, A. Van den Oord, and O. Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. *Adv. Neural Inform. Process. Syst.*, 32, 2019.

O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Med. Image Comput. Comput.-Assist. Interv.*, volume 9351 of *Lecture Notes Comput. Sci.*, pages 234–41. Springer, 2015.

J. Ruan, J. Li, and S. Xiang. VM-Unet: Vision mamba unet for medical image segmentation. *ACM Trans. on Multimedia Computing, Communications and Applications*, 2024.

D. Saxena and J. Cao. Generative adversarial networks (GANs) challenges, solutions, and future directions. *ACM Computing Surveys (CSUR)*, 54(3):1–42, 2021.

M. C. Schiappa, S. Azad, S. Vs, Y. Ge, O. Miksik, Y. S. Rawat, and V. Vineet. Robustness analysis on foundational segmentation models. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 1786–96. IEEE, 2024.

M. L. Seghier. Image segmentation evaluation with the dice index: Methodological issues, 2024.

Z. Shaaf, M. Jamil, R. Ambar, A. Alattab, A. Yahya, and Y. Asiri. Automatic left ventricle segmentation from short-axis cardiac MRI images based on fully convolutional neural network. *Diagnostics*, 12(2):414, 2022.

S. Shigwan, A. Gaikwad, and S. Awate. Object segmentation with deep neural nets coupled with a shape prior, when learning from a training set of limited quality and small size. In *IEEE Int. Symp. Biomed. Imaging*, pages 1149–53, 2020.

K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. *Adv. Neural Inform. Process. Syst.*, 28, 2015.

T. Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske Skrifter*, 5:1–34, 1948.

R. Strudel, R. Garcia, I. Laptev, and C. Schmid. Segmenter: Transformer for semantic segmentation. In *Int. Conf. Comput. Vis.*, pages 7262–72, 2021.

A. Suinesiaputra, B. Cowan, A. Al-Agamy, M. Elattar, N. Ayache, A. Fahmy, A. Khalifa, P. Medrano-Gracia, M. Jolly, A. Kadish, D. Lee, J. Margeta, S. Warfield, and A. Young. A collaborative resource to build consensus for automated left ventricular segmentation of cardiac MR images. *Med. Image Anal.*, 18(1):50–62, 2014.

F. Sun, Z. Luo, and S. Li. Boundary difference over union loss for medical image segmentation. In *Med. Image Comput. Comput.-Assist. Interv.*, volume 14223, pages 292–301, 2023.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2818–26, 2016.

H. Thisanke, C. Deshan, K. Chamith, S. Seneviratne, R. Vidanaarachchi, and D. Herath. Semantic segmentation using vision transformers: A survey. *Engineering Applications of Artificial Intelligence*, 126:106669, 2023.

B. Torpmann-Hagen, V. Thambawita, M. A. Riegler, P. Halvorsen, and K. Glette. Segmentation consistency training: out-of-distribution generalization for medical image segmentation. In *IEEE Int. Symposium on Multimedia*, pages 42–49. IEEE, 2022.

D. Tran, J. Snoek, and B. Lakshminarayanan. Practical uncertainty estimation and out-of-distribution robustness in deep learning. *Adv. Neural Inform. Process. Syst. Tutorial*, 2020.

T. L. van den Heuvel, D. de Bruijn, C. L. de Korte, and B. van Ginneken. Automated measurement of fetal head circumference using 2D ultrasound images. *PLoS ONE*, 13:e0200412, 2018.

A. van den Oord and O. Vinyals. Neural discrete representation learning. *Adv. Neural Inform. Process. Syst.*, 30, 2017.

H. Varma, A. V. Gaikwad, and S. P. Awate. Adversarial training with multiscale boundary-prediction DNN for robust topologically-constrained segmentation in ood images. In *IEEE Int. Symp. Biomed. Imaging*, pages 1–5. IEEE, 2023.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Adv. Neural Inform. Process. Syst.*, 30, 2017.

B. Wang, X. Gu, C. Fan, H. Xie, S. Zhang, X. Tian, and L. Gu. Sparse group composition for robust left ventricular epicardium segmentation. *Computerized Medical Imaging and Graphics*, 46:56–63, 2015.

Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–12, 2004.

R. Wen, H. Yuan, D. Ni, W. Xiao, and Y. Wu. From denoising training to test-time adaptation: Enhancing domain generalization for medical image segmentation. In *IEEE Winter Conf. Appl. Comput. Vis.*, pages 464–474, 2024.

J. Wu, R. Fu, H. Fang, Y. Zhang, Y. Yang, H. Xiong, H. Liu, and Y. Xu. MedSegDiff: Medical image segmentation with diffusion probabilistic model. In *Conf. Med. Imaging Deep Learn.*, volume 227 of *Proceedings of Machine Learning Research*, pages 1623–39, 2023.

J. Wu, W. Ji, H. Fu, M. Xu, Y. Jin, and Y. Xu. Medsegdiff-v2: Diffusion-based medical image segmentation with transformer. In *AAAI*, volume 38(6), pages 6030–38, 2024.

H. Xiao, L. Li, Q. Liu, X. Zhu, and Q. Zhang. Transformers in medical image segmentation: A review. *Biomed. Signal Process. Control.*, 84:104791, 2023.

Y. Xie, H. Chen, J. Qin, Y. Zhang, L. Dong, J. Du, T. Wang, and B. Lei. Towards semantically faithful diffusion representation for generalizable retinal image segmentation. *IEEE Trans. Med. Imag.*, 2025.

Z. Xing, T. Ye, Y. Yang, G. Liu, and L. Zhu. SegMamba: Long-range sequential modeling mamba for 3D medical image segmentation. In *Med. Image Comput. Comput.-Assist. Interv.*, pages 578–88. Springer, 2024.

Y. Xue, T. Xu, H. Zhang, L. Long, and X. Huang. SegAN: Adversarial network with multi-scale L1 loss for medical image segmentation. *Neuroinformatics*, 16:383–92, 2018.

S. Xun, D. Li, H. Zhu, M. Chen, J. Wang, J. Li, M. Chen, B. Wu, H. Zhang, X. Chai, Z. Jiang, Y. Zhang, and P. Huang. Generative adversarial networks in medical image segmentation: A review. *Comput. Biol. Med.*, 140: 105063, 2022.

W. Yan, Y. Wang, S. Gu, L. Huang, F. Yan, L. Xia, and Q. Tao. The domain shift problem of medical image segmentation and vendor-adaptation by UNet-GAN. In *Med. Image Comput. Comput.-Assist. Interv.*, pages 623–31. Springer, 2019.

V. Yeghiazaryan and I. Voiculescu. Family of boundary overlap metrics for the evaluation of medical image segmentation. *Journal of Medical Imaging*, 5(1), 2018.

W. Zeng, J. Luo, J. Cheng, and Y. Lu. Efficient fetal ultrasound image segmentation for automatic head circumference measurement using a lightweight deep convolutional neural network. *Medical Physics*, 49(8):5081–92, 2022.

Y. Zeng, P. Tsui, W. Wu, Z. Zhou, and S. Wu. Fetal ultrasound image segmentation for automatic head circumference biometry using deeply supervised attention-gated V-Net. *J. Digit. Imaging*, 34:134–48, 2021.

B. Zhang, Z. Tian, Q. Tang, X. Chu, X. Wei, C. Shen, and Y. Liu. SegVit: Semantic segmentation with plain vision transformers. *Adv. Neural Inform. Process. Syst.*, 35:4971–82, 2022.

H. Zhang, Y. Su, X. Xu, and K. Jia. Improving the generalization of segmentation foundation model under distribution shift via weakly supervised adaptation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 23385–95, 2024.

L. Zhang, F. Wu, K. Bronik, and B. W. Papiez. Diffuseg: domain-driven diffusion for medical image segmentation. *IEEE J. Biomed. Health Inform.*, 29(5):3619–31, 2025.

Z. Zhang, F. Yin, J. Liu, W. K. Wong, N. M. Tan, B. H. Lee, J. Cheng, and T. Y. Wong. ORIGA-light: An online retinal fundus image database for glaucoma analysis and research. In *IEEE Engineering in Medicine and Biology*, pages 3065–68, 2010.

Z. Zhang, Q. Liu, and Y. Wang. Road extraction by deep residual U-Net. *IEEE Geosci. Remote Sens. Lett.*, 15(5): 749–53, 2018.

H. Zhao, O. Gallo, I. Frosio, and J. Kautz. Loss functions for image restoration with neural networks. *IEEE Trans. Comput. Imaging*, 3(1):47–57, 2017.

Y. Zhao, J. Li, L. Ren, and Z. Chen. DTAN: Diffusion-based text attention network for medical image segmentation. *Comput. Biol. Med.*, 168:107728, 2024.

S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6881–90, 2021.

A. P. Zijdenbos, B. M. Dawant, R. A. Margolin, and A. C. Palmer. Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE Trans. Med. Imag.*, 13(4):716–724, 1994.