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

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

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

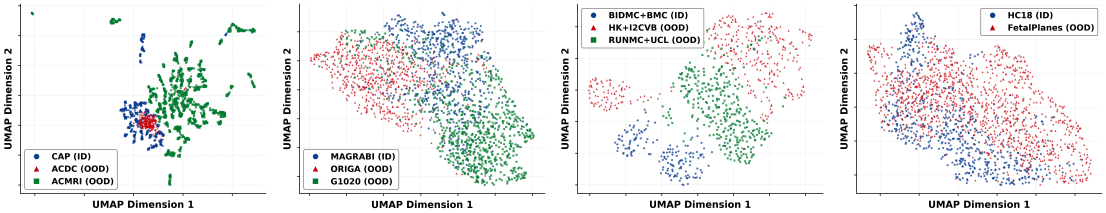
June 2026

MELBA Journal

Table of Contents

Agenda

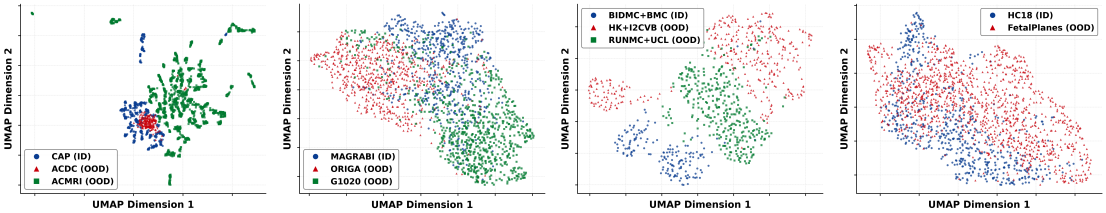
Introduction and Motivation



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

- DNNs suffer ***significant degradations when applied to out-of-distribution (OOD) data***.
- OOD images are ***same anatomical objects*** that are present in our training/in-dist (ID) data but ***acquired from different hospitals/scanners*** - leading to ***distribution shifts***.

Introduction and Motivation



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

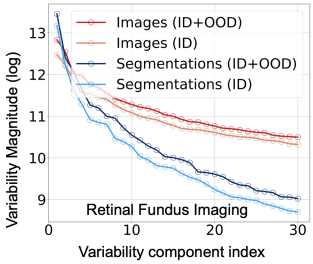
- DNNs suffer ***significant degradations when applied to out-of-distribution (OOD) data***.
- OOD images are ***same anatomical objects*** that are present in our training/in-dist (ID) data but ***acquired from different hospitals/scanners*** - leading to ***distribution shifts***.
- Our novel VarDeepPCA, a small, lightweight framework corrects degraded o/p of existing DNNs on OOD data by ***leveraging priors on the inherent geometry*** of the object of interest which ***remains highly invariant*** to the aforementioned OOD variations.
- The low-dimensional distribution of valid anatomical geometry serves as a powerful prior to ***rectify the erroneous segmentation*** produced by existing DNNs on OOD images.

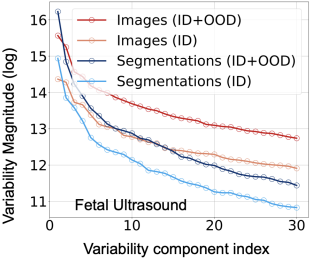
Introduction and Motivation

- Unlike learning from target-domain data or extreme pretraining, our VarDeepPCA explicitly learns a distribution of valid anatomical geometries using ***small ID datasets ($\sim$ 150 samples)*** — also giving uncertainty estimates.
- **Myocardial segmentation** in short-axis cardiac MRI is essential for estimating contractility and tissue strain — which aids in ***diagnosing infraction, ischemia and ventricular dyssynchrony***.

Introduction and Motivation

- Unlike learning from target-domain data or extreme pretraining, our VarDeepPCA explicitly learns a distribution of valid anatomical geometries using ***small ID datasets ($\sim$ 150 samples)*** — also giving uncertainty estimates.
- **Myocardial segmentation** in short-axis cardiac MRI is essential for estimating contractility and tissue strain — which aids in ***diagnosing infraction, ischemia and ventricular dyssynchrony***.
- In **ophthalmology**, the optic disc and cup segmentation from the retinal scans, i.e., the neuroretinal rim enables the calculation of the cup-to-disc ratio — a key ***biomarker for monitoring glaucoma***.
- **Prostate segmentation** in T2-weighted MRI is pivotal for diagnosis, staging and treatment planning of ***prostate cancer***.
- **Fetal head segmentation** in ultrasound images facilitates the measurement of geometric parameters to ***assess fetal growth and detect developmental anomalies***.





Agenda

Methodology: VarDeepPCA and the reinterpretation of Softmax

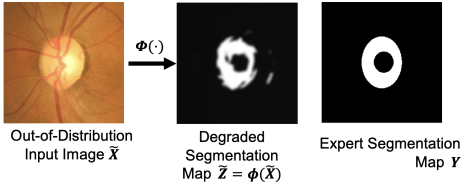
- $X :=$ medical image; $Y :=$ ground truth segmentation map; $\Phi(\cdot) :=$ existing DNN segmenter; $\widetilde{Z} := \Phi(\widetilde{X})$ anatomically implausible segmentation.
- Correct $\widetilde{Z}$ by training a ***lightweight encoder decoder model*** (VarDeepPCA) on the same segmentation maps that were used to train $\Phi(\cdot)$, and without access to any intensity images.

Methodology: VarDeepPCA and the reinterpretation of Softmax

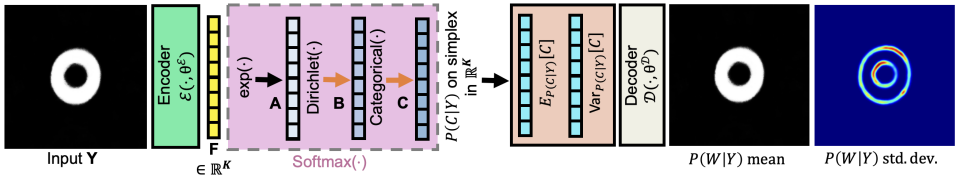
- $X :=$ medical image; $Y :=$ ground truth segmentation map; $\Phi(\cdot) :=$ existing DNN segmenter; $\widetilde{Z} := \Phi(\widetilde{X})$ anatomically implausible segmentation.
- Correct $\widetilde{Z}$ by training a ***lightweight encoder decoder model*** (VarDeepPCA) on the same segmentation maps that were used to train $\Phi(\cdot)$, and without access to any intensity images.
- VarDeepPCA's encoder maps the input Y segmentation map to $F := \mathcal{E}(Y; \theta^{\mathcal{E}})$.
- We hypothesize the variability of a given segmentation map can be captured by K principal (non-linear) modes of variation, i.e., $C = \mathbb{1}_k$, given by $P(C|Y)$.
- A segmentation map lies in a $K-1$ dimensional simplex.
- $[P(C = \mathbb{1}_1|Y), \ldots, P(C = \mathbb{1}_K|Y)]$ defines the association of Y with each of the K modes.

Existing DNN Segmenters performs well on ID images

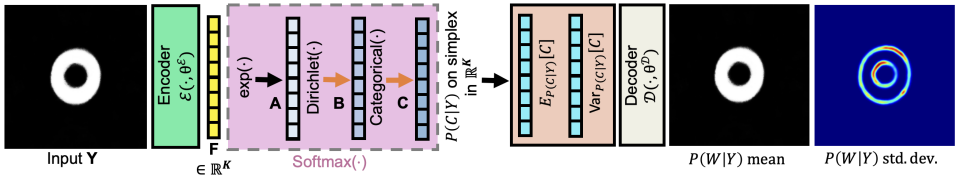
Existing DNN Segmenters fail on OOD images



Methodology: VarDeepPCA and the reinterpretation of Softmax

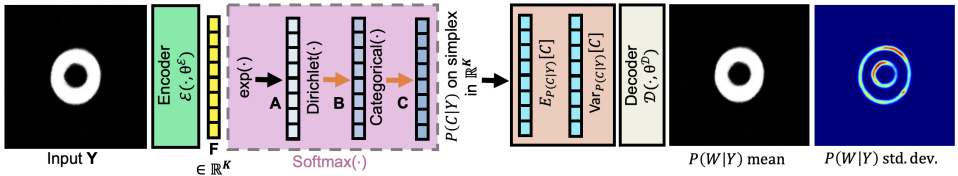


Methodology: VarDeepPCA and the reinterpretation of Softmax



- VarDeepPCA reinterprets the softmax mapping in a variational setup $P(C|Y) \equiv \mathrm{Softmax}(F)$.
- We model $P(C|F)$ as the *posterior-predictive distribution* on C arising from:
  - *Categorical*-distribution *likelihood* $P(C|\cdot)$ on the modes of variation indicated by C, and

Methodology: VarDeepPCA and the reinterpretation of Softmax



- VarDeepPCA reinterprets the softmax mapping in a variational setup $P(C|Y) \equiv \text{Softmax}(F)$.
- We model $P(C|F)$ as the *posterior-predictive distribution* on C arising from:
  - *Categorical*-distribution *likelihood* $P(C|\cdot)$ on the modes of variation indicated by C, and
  - A *Dirichlet*-distribution (conjugate) *prior* $P(\cdot|F)$.
- Consider a random vector A having elements, $A_k := \exp(F_k) > 0 \;\forall\; 1 \leq k \leq K$.
- A parameterizes a Dirichlet distribution $\text{Dir}(B; A)$ of a hidden random vector B residing on the $(K-1)$-dimensional simplex.
- $Y \to F \to A$ is deterministic, hence the following equivalence between posterior-predictive distributions holds: $P(C|Y=y) \equiv P(C|F=\mathcal{E}(y;\theta^{\mathcal{E}})) \equiv P(C|A=\exp(\mathcal{E}(y;\theta^{\mathcal{E}})))$.

Methodology: VarDeepPCA and the reinterpretation of Softmax

- Consider a categorical distribution $\text{Cat}(C; B)$ on one-hot vectors C, parameterized by the hidden random vector B that is sampled from 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

$$\int_{\boldsymbol{b}} \mathbf{Cat}(\boldsymbol{C}; \boldsymbol{b})\mathbf{Dir}(\boldsymbol{b}; \boldsymbol{A})d\boldsymbol{b} \tag{2}$$

Methodology: VarDeepPCA and the reinterpretation of Softmax

- Consider a categorical distribution $\text{Cat}(C; B)$ on one-hot vectors C, parameterized by the hidden random vector B that is sampled from 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

$$\int_{\boldsymbol{b}} \textbf{Cat}(\boldsymbol{C}; \boldsymbol{b})\textbf{Dir}(\boldsymbol{b}; \boldsymbol{A})\boldsymbol{db} \tag{2}$$

which equals

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

- 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.

Methodology: VarDeepPCA and the reinterpretation of Softmax

- This yields

$$P(C|A) = \frac{1}{\eta(A)} \int_b \prod_k (b_k)^{C_k + A_k - 1} db = \frac{\boldsymbol{\eta(A + C)}}{\boldsymbol{\eta(A)}} = \frac{\prod_k \Gamma(A_k + C_k) / \Gamma\left(\sum_k (A_k + C_k)\right)}{\prod_k \Gamma(A_k) / \Gamma(\sum_k A_k)}. \tag{4}$$

- Because C is a one-hot vector, $\sum_k C_k = 1$.

Methodology: VarDeepPCA and the reinterpretation of Softmax

- This yields

$$P(C|A) = \frac{1}{\eta(A)} \int_b \prod_k (b_k)^{C_k + A_k - 1} db = \frac{\boldsymbol{\eta(A + C)}}{\boldsymbol{\eta(A)}} = \frac{\prod_k \Gamma(A_k + C_k) / \Gamma\left(\sum_k (A_k + C_k)\right)}{\prod_k \Gamma(A_k) / \Gamma(\sum_k A_k)}. \tag{4}$$

- 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

$$P(C = \mathbb{1}_k|A) = \frac{\Gamma(A_k + 1) \prod_{j \neq k} \Gamma(A_j) / \Gamma(\sum_j A_j + 1)}{\prod_j \Gamma(A_j) / \Gamma(\sum_j A_j)} = \frac{A_k \Gamma(A_k) \prod_{j \neq k} \Gamma(A_j)}{\prod_j \Gamma(A_j)} \cdot \frac{\Gamma(\sum_j A_j)}{(\sum_j A_j)\Gamma(\sum_j A_j)}$$

$$= \frac{A_k}{\sum_{j=1}^K A_j} = \frac{\boldsymbol{\exp(F_k)}}{\sum_{\boldsymbol{j=1}}^{\boldsymbol{K}} \boldsymbol{\exp(F_j)}} \tag{5}$$

Methodology: VarDeepPCA and the reinterpretation of Softmax

- This yields

$$P(C|A) = \frac{1}{\eta(A)} \int_b \prod_k (b_k)^{C_k + A_k - 1} db = \frac{\boldsymbol{\eta(A + C)}}{\boldsymbol{\eta(A)}} = \frac{\prod_k \Gamma(A_k + C_k) / \Gamma\left(\sum_k (A_k + C_k)\right)}{\prod_k \Gamma(A_k) / \Gamma(\sum_k A_k)}. \tag{4}$$

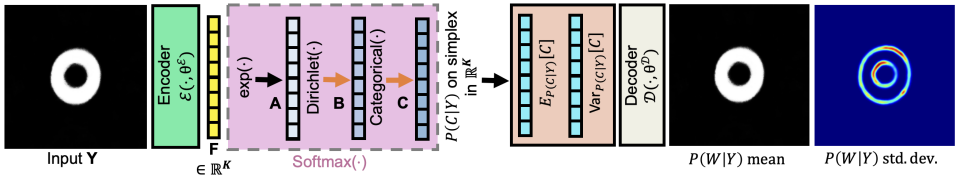
- Because C is a one-hot vector, $\sum_k C_k = 1$.
- Using the property $\Gamma(g+1) = g\Gamma(g)$, for gamma functions, we simplify the posterior-predictive distribution as

$$P(C = \mathbb{1}_k|A) = \frac{\Gamma(A_k + 1) \prod_{j \neq k} \Gamma(A_j) / \Gamma(\sum_j A_j + 1)}{\prod_j \Gamma(A_j) / \Gamma(\sum_j A_j)} = \frac{A_k \Gamma(A_k) \prod_{j \neq k} \Gamma(A_j)}{\prod_j \Gamma(A_j)} \cdot \frac{\Gamma(\sum_j A_j)}{(\sum_j A_j)\Gamma(\sum_j A_j)}$$

$$= \frac{A_k}{\sum_{j=1}^K A_j} = \frac{\boldsymbol{\exp(F_k)}}{\sum_{\boldsymbol{j=1}}^{\boldsymbol{K}} \boldsymbol{\exp(F_j)}} \tag{5}$$

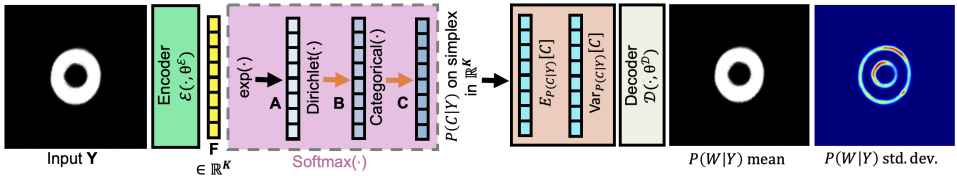
- Softmax mapping $F \to P(C|Y)$ is deterministic, it subsumes variational modeling by defining prior and the likelihood distribution, i.e., $P(B|\exp(F)) \equiv P(B|A)$ and $P(C|B)$.
- Marginalizes out the random variable B via Bayesian inference to produce the *analytically exact* posterior predictive distribution $P(C|F)$ in *closed form*.

Methodology: VarDeepPCA and the reinterpretation of Softmax



- For a given input Y, the latent distribution $P(C|Y)$ is produced by variational modeling by implicitly modeling the Categorical distribution $P(C|B)$ and the Dirichlet distribution $P(B|Y = y) \equiv P(B|A = \exp(\mathcal{E}(y; \theta^{\mathcal{E}}))$.

Methodology: VarDeepPCA and the reinterpretation of Softmax



- For a given input Y, the latent distribution $P(C|Y)$ is produced by variational modeling 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:
  - map input Y to $F \leftarrow \mathcal{E}(Y;\theta^{\mathcal{E}})$ and then map F to $A \leftarrow \exp(F)$
  - sample $b \sim \mathrm{Dir}(B;A)$ and then sample $c \sim \mathrm{Cat}(C;b)$
- The mean and variance are available analytically in closed form for the posterior predictive categorical distribution $P(C|Y)$:

$$C^{\mathsf{mean}} := \mathbb{E}_{P(C|Y;\theta^{\mathcal{E}})}[C] = [P(C=\mathbb{1}_1|Y), \ldots, P(C=\mathbb{1}_K|Y)] = \mathsf{Softmax}(\mathcal{E}(Y;\theta^{\mathcal{E}})) \quad (6)$$

$$C_k^{\mathsf{var}} := C_k^{\mathsf{mean}}(1 - C_k^{\mathsf{mean}}). \quad (7)$$

Methodology: Learning formulation and Uncertainty Estimation

- 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,

$$\boldsymbol{M} := \boldsymbol{\mathcal{D}}(\boldsymbol{C}^{\text{mean}}; \boldsymbol{\theta}^{\boldsymbol{\mathcal{D}}}), \text{ and} \tag{8}$$

$$\boldsymbol{V_i} := \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{C_k^{\text{var}}} \left(\frac{\boldsymbol{\partial \mathcal{D}_i(L)}}{\boldsymbol{\partial L_k}} \bigg|_{\boldsymbol{L:=C^{\text{mean}}=\text{Softmax}(\mathcal{E}(Y;\theta^{\mathcal{E}}))}} \right)^{\boldsymbol{2}}. \tag{9}$$

Methodology: Learning formulation and Uncertainty Estimation

- 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,

$$\boldsymbol{M} := \boldsymbol{\mathcal{D}}(\boldsymbol{C}^{\text{mean}}; \boldsymbol{\theta}^{\boldsymbol{\mathcal{D}}}), \text{ and} \tag{8}$$

$$\boldsymbol{V_i} := \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{C_k^{\text{var}}} \left(\frac{\boldsymbol{\partial \mathcal{D}_i(L)}}{\boldsymbol{\partial L_k}} \bigg|_{\boldsymbol{L:=C^{\text{mean}}=\text{Softmax}(\mathcal{E}(Y;\theta^{\mathcal{E}}))}} \right)^{\boldsymbol{2}}. \tag{9}$$

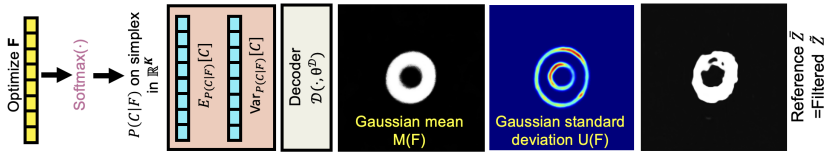
- We formulate the variational learning objective to maximize, over parameters θ:

$$\arg\max_{\theta} \prod_{n=1}^{N} \mathcal{N}(Y_n; M(Y_n;\theta), V(Y_n;\theta)) \equiv \tag{10}$$

$$\boldsymbol{\arg\min_{\theta}} \sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{I}} \frac{\boldsymbol{(Y_{ni} - M_i(Y_n;\theta))^2}}{\boldsymbol{V_i(Y_n;\theta) + \epsilon}} + \boldsymbol{\log(V_i(Y_n;\theta) + \epsilon)}, \tag{11}$$

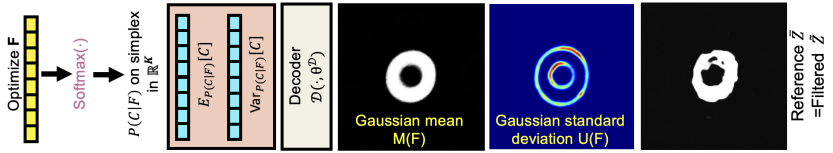
- Despite modelling (i) a latent distribution $P(C|Y)$ and (ii) distributions $P(C|B)$ and $P(B|Y)$ implicitly within the softmax parameterization, VarDeepPCA eliminates the need for Monte Carlo sampling and the associated reparameterization.

Methodology: Reviving poor segmentation map using VarDeepPCA



- We propose a novel two-stage algorithm for restoring the poor segmentation maps $\widetilde{Z}$.
- 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 $\overline{Z} := \mathcal{D}(\mathrm{Softmax}(\mathcal{E}(\widetilde{Z}; \theta^{\mathcal{E}})); \theta^{\mathcal{D}})$.

Methodology: Reviving poor segmentation map using VarDeepPCA



- We propose a novel two-stage algorithm for restoring the poor segmentation maps $\widetilde{Z}$.
- 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 $\overline{Z} := \mathcal{D}(\text{Softmax}(\mathcal{E}(\widetilde{Z}; \theta^{\mathcal{E}})); \theta^{\mathcal{D}})$.
- 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
  $\boldsymbol{F^* := \arg\max_F \mathcal{N}(\overline{Z}; M(F; \theta^{\mathcal{D}}), V(F; \theta^{\mathcal{D}}))}$ using gradient ascent, and
  (iii) obtaining the restored segmentation $M^* := \mathcal{D}(\text{Softmax}(F^*); \theta^{\mathcal{D}})$ with the associated per-pixel uncertainties given by $U_i^* := \sqrt{V_i^*}$.

Agenda

Datasets and Data augmentations

Application (Modality)	**ID Dataset**	**ID Train**	**ID Val**	**ID Test**	**OOD Dataset 1 (Test Size)**	**OOD Dataset 2 (Test Size)**
Myocardium (MRI)	CAP [1–3]	**150**	**70**	**634**	ACDC [4] **(220 samples)**	ACMRI [5] **(1722 samples)**
Neuroretinal Rim (Fundus)	Magrabi [6]	**150**	**63**	**620**	ORIGA [7] **(637 samples)**	G1020 [8] **(788 samples)**
Prostate (MRI)	BIDMC+BMC [9–11]	**150**	**18**	**213**	HK+I2CVB [9, 12] **(366 samples)**	RUNMC+UCL [10, 11, 9] **(348 samples)**
Fetal Head (Ultrasound)	HC18 [13]	**150**	**68**	**666**	FetalPlanes [14] **(1250 samples)**	–

- Even with 150 samples, early DNNs like UNet are generalizing well to ID test set, having $\sim$90% Dice Coefficient and HD95$<$6.
- Fundamental limitation of DNNs — **not generalizing to OOD datasets** of the same anatomy.

Datasets and Data augmentations

Application (Modality)	ID Dataset	ID Train	ID Val	ID Test	OOD Dataset 1 (Test Size)	OOD Dataset 2 (Test Size)
Myocardium (MRI)	CAP [1–3]	**150**	**70**	**634**	ACDC [4] **(220 samples)**	ACMRI [5] **(1722 samples)**
Neuroretinal Rim (Fundus)	Magrabi [6]	**150**	**63**	**620**	ORIGA [7] **(637 samples)**	G1020 [8] **(788 samples)**
Prostate (MRI)	BIDMC+BMC [9–11]	**150**	**18**	**213**	HK+I2CVB [9, 12] **(366 samples)**	RUNMC+UCL [10, 11, 9] **(348 samples)**
Fetal Head (Ultrasound)	HC18 [13]	**150**	**68**	**666**	FetalPlanes [14] **(1250 samples)**	–

- Even with 150 samples, early DNNs like UNet are generalizing well to ID test set, having $\sim$90% Dice Coefficient and HD95$<$6.
- Fundamental limitation of DNNs — **not generalizing to OOD datasets** of the same anatomy.
- **Data augmentation:** applied across all models to enhance generalization and mitigate overfitting.
- **Geometric Transforms:** Synchronous horizontal/vertical flips and affine transforms (rotation, scaling, translation) applied to images and segmentation maps.
- **Pixel-Level Transforms:** Random brightness-contrast adjustments, random gamma, and blur applied to input images.

Results (Quantitative): Myocardium Segmentation

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

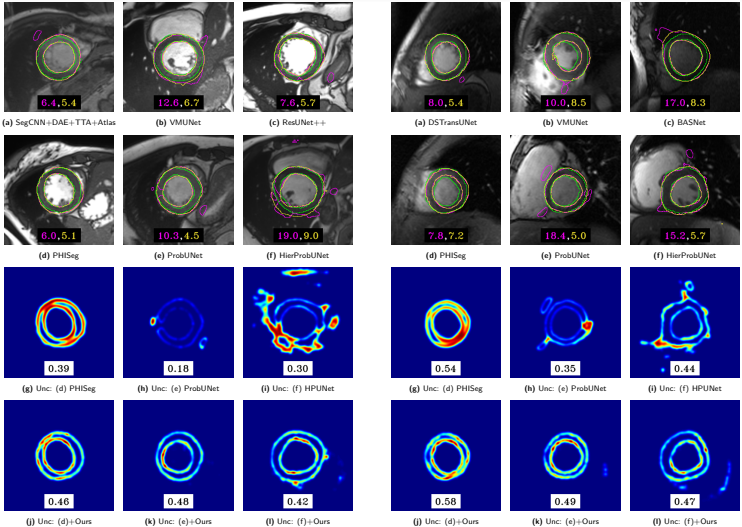
Results (Quantitative): Myocardium Segmentation

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

- Results are presented as mean ± standard deviation of the respective metrics.
- Statistically significant improvements based on a two-tailed paired sample t-test (**bold**, $p < 0.05$).
- VarDeepPCA results are either as good or better than the baselines; Baseline improvements: 8/15 for CAP (ID), 10/15 for ACDC (OOD), and 12/15 for ACMRI (OOD).

Results (Qualitative): Myocardium Segmentation

- Top three baselines on respective OOD dataset based on HD95 performance are selected for comparison.
- Segmentation maps produced by variational/probabilistic baselines, i.e., PHISeg, ProbUNet and HierProbUNet are shown along with their HD95 values.
- Normalized cross correlation between the error maps and the uncertainty maps are presented, showing VarDeepPCA improves uncertainty calibration over probabilistic baselines.



Results (Quantitative): Myocardium Uncertainty Quantification

Models	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	.53 ± .04	.83 ± .03	.23 ± .05	**.58 ± .03**	**.85 ± .02**	**.18 ± .03**	.41 ± .05	.73 ± .08	.35 ± .08	**.48 ± .05**	**.88 ± .07**	**.16 ± .03**	.52 ± .04	.75 ± .07	.32 ± .06	**.56 ± .05**	**.85 ± .06**	**.10 ± .02**
ProbUNet	.34 ± .06	.75 ± .01	.29 ± .06	**.50 ± .03**	**.91 ± .01**	**.19 ± .03**	.23 ± .05	.63 ± .04	.31 ± .12	**.46 ± .04**	**.89 ± .04**	**.21 ± .03**	.33 ± .08	.73 ± .04	.34 ± .10	**.45 ± .04**	**.87 ± .05**	**.12 ± .02**
HPUNet	.45 ± .03	.73 ± .02	.18 ± .07	**.50 ± .03**	**.85 ± .02**	**.10 ± .03**	.34 ± .04	.61 ± .04	.49 ± .07	**.43 ± .05**	**.75 ± .04**	**.26 ± .02**	.42 ± .05	.57 ± .05	.31 ± .09	**.47 ± .04**	**.82 ± .04**	**.19 ± .02**

- We measure the normalized cross correlation (NCC), unified score (US) and thresholded adaptive calibration error (TACE).
- Results are presented as mean $\pm$ standard deviation of the respective metrics.
- Augmenting existing DNNs with VarDeepPCA plugins yields statistically significant improvements based on a two-tailed paired sample t-test (**bold**, $p < 0.05$).

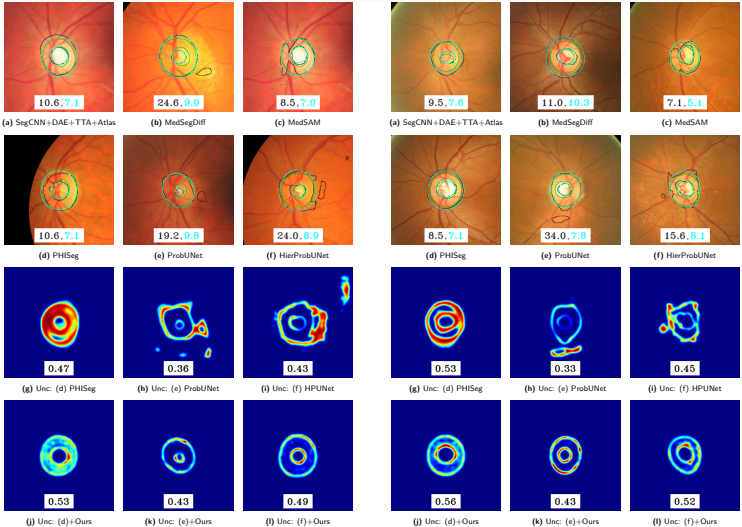
Results (Quantitative): Neuroretinal Rim Segmentation

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

- Results are presented as mean ± standard deviation of the respective metrics.
- Statistically significant improvements based on a two-tailed paired sample t-test (**bold**, $p < 0.05$).
- Baseline improvements: 11/15 for MAGRABI (ID), 14/15 for G1020 (OOD), and 13/15 for ORIGA (OOD).

Results (Qualitative): Neuroretinal Rim Segmentation

- Top three baselines on respective OOD dataset based on HD95 performance are selected for comparison.
- Segmentation maps produced by variational/probabilistic baselines, i.e., PHISeg, ProbUNet and HierProbUNet are shown along with their HD95 values.
- Normalized cross correlation between the error maps and the uncertainty maps are presented, showing VarDeepPCA improves uncertainty calibration over probabilistic baselines.



Results (Quantitative): Neuroretinal Rim Uncertainty Quantification

Models	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	.56 ± .05	.93 ± .03	.21 ± .04	**.58 ± .04**	**.95 ± .02**	**.12 ± .01**	.48 ± .07	.82 ± .11	.32 ± .06	**.57 ± .05**	**.87 ± .06**	**.13 ± .01**	.54 ± .05	.82 ± .04	.48 ± .06	**.56 ± .04**	**.86 ± .03**	**.18 ± .01**
ProbUNet	.42 ± .07	.91 ± .04	.27 ± .08	**.49 ± .06**	**.94 ± .03**	**.12 ± .02**	.37 ± .08	.85 ± .06	.39 ± .12	**.43 ± .07**	**.88 ± .06**	**.13 ± .01**	.33 ± .08	.86 ± .07	.55 ± .11	**.39 ± .05**	**.88 ± .06**	**.16 ± .01**
HPUNet	.45 ± .04	.92 ± .03	.19 ± .09	**.53 ± .05**	**.95 ± .02**	**.10 ± .01**	.44 ± .07	.88 ± .06	.40 ± .12	**.49 ± .07**	**.92 ± .04**	**.13 ± .01**	.43 ± .05	.88 ± .04	.45 ± .09	**.53 ± .05**	**.93 ± .04**	**.12 ± .01**

- We measure the normalized cross correlation (NCC), unified score (US) and thresholded adaptive calibration error (TACE).
- Results are presented as mean ± standard deviation of the respective metrics.
- Augmenting existing DNNs with VarDeepPCA plugins yields statistically significant improvements based on a two-tailed paired sample t-test (**bold**, $p < 0.05$).

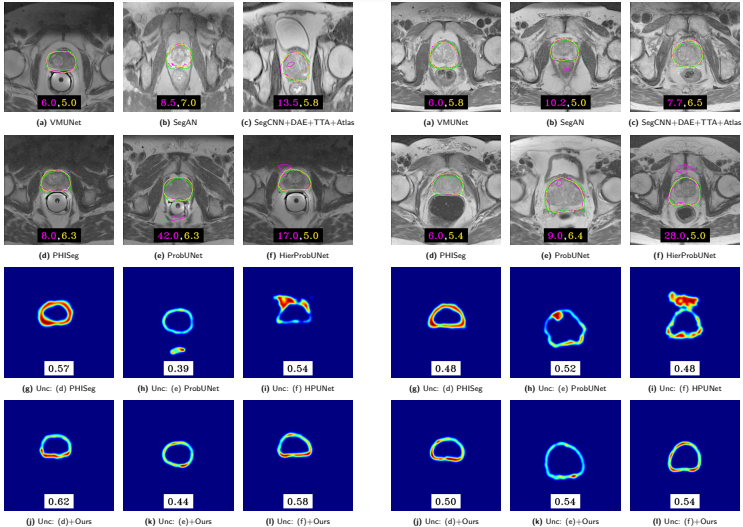
Results (Quantitative): Prostate Segmentation

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

- Results are presented as mean ± standard deviation of the respective metrics.
- Statistically significant improvements based on a two-tailed paired sample t-test (**bold**, $p < 0.05$).
- Baseline improvements: 9/15 on BIDMC+BMC (ID) on HD95, 12/15 on HK+I2CVB (OOD), and 11/15 on RUNMC+UCL (OOD) dataset.

Results (Qualitative): Prostate Segmentation

- Top three baselines on respective OOD dataset based on HD95 performance are selected for comparison.
- Segmentation maps produced by variational/probabilistic baselines, i.e., PHISeg, ProbUNet and HierProbUNet are shown along with their HD95 values.
- Normalized cross correlation between the error maps and the uncertainty maps are presented, showing VarDeepPCA improves uncertainty calibration over probabilistic baselines.



Results (Quantitative): Prostate Uncertainty Quantification

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	.66 ± .14	.76 ± .04	.39 ± .26	**.69 ± .10**	**.82 ± .04**	**.12 ± .24**	.57 ± .11	.63 ± .03	.47 ± .23	**.63 ± .08**	**.69 ± .04**	**.14 ± .24**	.45 ± .05	.66 ± .04	.53 ± .17	**.52 ± .03**	**.73 ± .04**	**.09 ± .12**
ProbUNet	.45 ± .08	.69 ± .03	.44 ± .31	**.54 ± .08**	**.73 ± .04**	**.14 ± .25**	.40 ± .06	.59 ± .03	.50 ± .32	**.45 ± .06**	**.65 ± .03**	**.12 ± .23**	.51 ± .03	.68 ± .03	.46 ± .25	**.55 ± .02**	**.74 ± .03**	**.08 ± .12**
HPUNet	.51 ± .09	.69 ± .03	.43 ± .29	**.61 ± .08**	**.79 ± .03**	**.16 ± .24**	.54 ± .07	.64 ± .03	.65 ± .25	**.59 ± .06**	**.75 ± .02**	**.19 ± .18**	.47 ± .04	.65 ± .04	.46 ± .22	**.56 ± .03**	**.74 ± .03**	**.12 ± .15**

- We measure the normalized cross correlation (NCC), unified score (US) and thresholded adaptive calibration error (TACE).
- Results are presented as mean ± standard deviation of the respective metrics.
- Augmenting existing DNNs with VarDeepPCA plugins yields statistically significant improvements based on a two-tailed paired sample t-test (**bold**, $p < 0.05$).

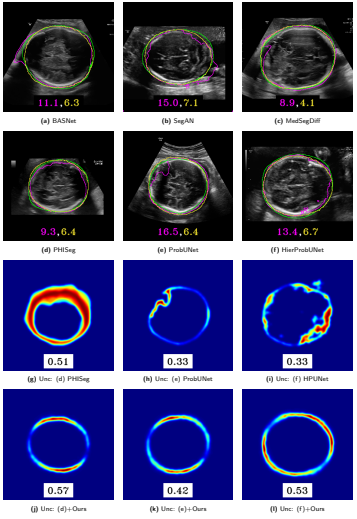
Results (Quantitative): Fetal Head Segmentation

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

- Results are presented as mean ± standard deviation of the respective metrics.
- Statistically significant improvements based on a two-tailed paired sample t-test (**bold**, $p < 0.05$).
- Baseline improvements: 15/15 for both HC18 (ID) and FetalPlanes (OOD) dataset.

Results (Qualitative): Fetal Head Segmentation

- Top three baselines on respective OOD dataset based on HD95 performance are selected for comparison.
- Segmentation maps produced by variational/probabilistic baselines, i.e., PHISeg, ProbUNet and HierProbUNet are shown along with their HD95 values.
- Normalized cross correlation between the error maps and the uncertainty maps are presented, showing VarDeepPCA improves uncertainty calibration over probabilistic baselines.



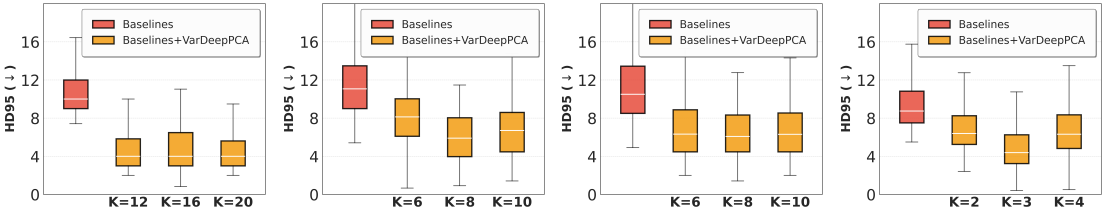
FetalPlanes (OOD Test Set)

Results (Quantitative): Fetal Head Uncertainty Quantification

Models	HC18 (ID)						FetalPlanes (OOD)					
	Baseline			Baseline + **VarDeepPCA**			Baseline			Baseline + **VarDeepPCA**		
	NCC ↑	US ↑	TACE ↓	NCC ↑	US ↑	TACE ↓	NCC ↑	US ↑	TACE ↓	NCC ↑	US ↑	TACE ↓
PHISeg	.67 ± .06	.69 ± .03	.34 ± .04	**.73 ± .03**	**.75 ± .03**	**.14 ± .03**	.49 ± .05	.71 ± .02	.45 ± .05	**.55 ± .04**	**.78 ± .00**	**.08 ± .04**
ProbUNet	.54 ± .05	.72 ± .03	.37 ± .08	**.68 ± .03**	**.82 ± .04**	**.15 ± .04**	.34 ± .07	.65 ± .02	.44 ± .03	**.45 ± .05**	**.75 ± .01**	**.11 ± .03**
HPUNet	.46 ± .06	.71 ± .02	.41 ± .04	**.54 ± .03**	**.77 ± .03**	**.25 ± .03**	.35 ± .06	.65 ± .01	.46 ± .04	**.53 ± .05**	**.74 ± .00**	**.14 ± .03**

- We measure the normalized cross correlation (NCC), unified score (US) and thresholded adaptive calibration error (TACE).
- Results are presented as mean ± standard deviation of the respective metrics.
- Augmenting existing DNNs with VarDeepPCA plugins yields statistically significant improvements based on a two-tailed paired sample t-test (**bold**, $p < 0.05$).

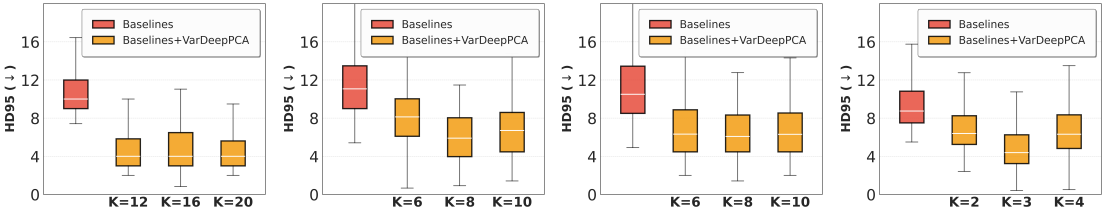
Sensitivity analysis: Changing latent dimension K with VarDeepPCA



(a) Myocardium **(b) Neuroretinal Rim** **(c) Prostate** **(d) Fetal Head**

- Selected VMUNet, PHISeg (best performing) and UNet (as a representative of the early DNNs).
- We select CAP (ID) + ACDC (OOD) for myocardium; MAGRABI (ID) + ORIGA (OOD) for neuroretinal rim; BIDMC+BMC (ID)+HK+I2CVB (OOD) for prostate; HC18 (ID) + FetalPlanes (OOD) for fetal head segmentation.

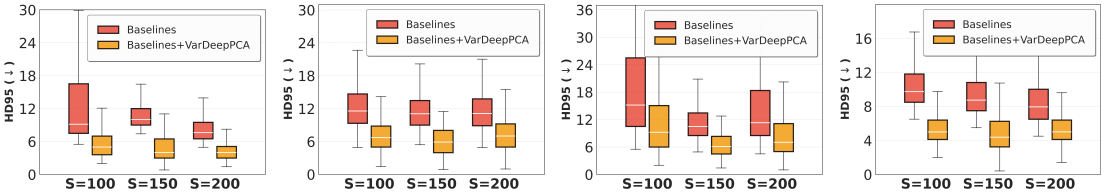
Sensitivity analysis: Changing latent dimension K with VarDeepPCA



(a) Myocardium **(b) Neuroretinal Rim** **(c) Prostate** **(d) Fetal Head**

- Selected VMUNet, PHISeg (best performing) and UNet (as a representative of the early DNNs).
- We select CAP (ID) + ACDC (OOD) for myocardium; MAGRABI (ID) + ORIGA (OOD) for neuroretinal rim; BIDMC+BMC (ID)+HK+I2CVB (OOD) for prostate; HC18 (ID) + FetalPlanes (OOD) for fetal head segmentation.
- Optimized K for each application based on validation dataset during training.
- VarDeepPCA consistently improved OOD performance.

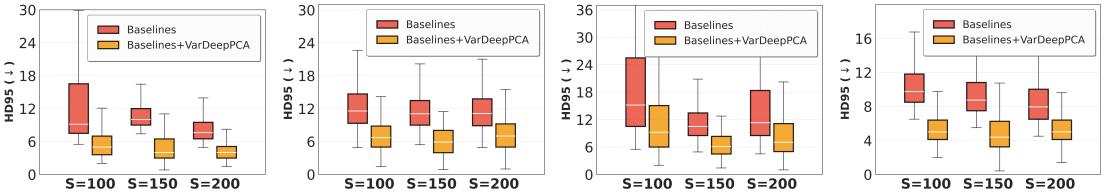
Sensitivity analysis: changing size of the training set S



(e) Myocardium (f) Neuroretinal Rim (g) Prostate (h) Fetal Head

- We selected the same baselines and the same datasets (i.e., one ID and one OOD) for this analysis.
- Used $S_{100} \subset S_{150} \subset S_{200}$ as training samples sizes.

Sensitivity analysis: changing size of the training set S



(e) Myocardium (f) Neuroretinal Rim (g) Prostate (h) Fetal Head

- We selected the same baselines and the same datasets (i.e., one ID and one OOD) for this analysis.
- Used $S_{100} \subset S_{150} \subset S_{200}$ as training samples sizes.
- We find the performance of the methods to be quite insensitive to small changes in S.
- Across all training set and sample sizes, our plugin consistently and significantly improved over the baselines.

Agenda

Conclusion and Summary of Contributions

- Presented VarDeepPCA - novel framework/plugin to ***restore degraded segmentation produced by DNNs on OOD images***.
- Learns principal modes of anatomical geometric variations from a ***small ID dataset of segmentation maps***.
- Leverages a ***reinterpretation of softmax mapping to implicitly perform exact distribution modelling***, thereby enabling computationally efficient sampling free learning and inference.

Conclusion and Summary of Contributions

- Presented VarDeepPCA - novel framework/plugin to ***restore degraded segmentation produced by DNNs on OOD images***.
- Learns principal modes of anatomical geometric variations from a ***small ID dataset of segmentation maps***.
- Leverages a ***reinterpretation of softmax mapping to implicitly perform exact distribution modelling***, thereby enabling computationally efficient sampling free learning and inference.
- VarDeepPCA ***doesn't require access to any OOD data nor to any medical intensity image***.
- Results across ***14 publicly available datasets*** and ***15 DNN segmenters*** show consistent improvement of DNNs in OOD segmentation, boundary delineation accuracy, plausability of anatomical geometry, and uncertainty estimates.

Constraints of our framework and Future work

- Our method relies on ***consistency of geometry of anatomical object of interest across dataset*** — inherently unsuitable for segmenting pathologies with highly non-uniform/amorphous/unpredictable shapes/geometries like lesions, tumors etc.
- If a segmentation map lies within the learned distribution of valid geometries but is still incorrect, then VarDeepPCA is unable to correct/improve the segmentation.

Constraints of our framework and Future work

- Our method relies on ***consistency of geometry of anatomical object of interest across dataset*** — inherently unsuitable for segmenting pathologies with highly non-uniform/amorphous/unpredictable shapes/geometries like lesions, tumors etc.
- If a segmentation map lies within the learned distribution of valid geometries but is still incorrect, then VarDeepPCA is unable to correct/improve the segmentation.
- Extreme errors in the segmentation produced by a DNN — then our/any other method will be unable to restore the correct segmentation — even in such cases VarDeepPCA succeeds in projecting the segmentation to one of a valid anatomical geometry.
- ***Future work:*** Multiclass segmentation problem in 2D and 3D images.

References I

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

References II

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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, 2018.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

References III

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] J. Clough, N. Byrne, I. Oksuz, V. Zimmer, J. Schnabel, and A. King. A topological loss function for deep-learning based image segmentation using persistent homology. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12):8766–78, 2022.

[31] 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.

[32] A. Gaikwad, H. Varma, and S. 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, 2023.

[33] W. Zheng, J. Chen, K. Zhang, J. Yan, J. Wang, Y. Cheng, B. Du, D. Z. Chen, H. Gao, J. Wu, and H. Xu. Polygonal approximation learning for convex object segmentation in biomedical images with bounding box supervision. *IEEE J. Biomed. Health Inform.*, 28(8):4522–33, 2023.

Any Questions?

Thank You!

pal.jimut@iitb.ac.in