On the Out of Distribution Robustness of Foundation Models in Medical Image Segmentation

Duy M. H. Nguyen*[1,2,3], Tan N. Pham*[3,4], Nghiem T. Diep[3], Nghi Phan[3], Quang Pham[5], Vinh Tong[1,2], Binh T. Nguyen[4], Ngan Le[6], Nhat Ho[7], Pengtao Xie[8,9], Daniel Sonntag[3,10], Mathias Niepert[1,2]

[1]University of Stuttgart, [2]IMPRS for Intelligent Systems,
[3]German Research Center for Artificial Intelligence, [4]University of Science - VNUHCM,
[5]Singapore Management University, [6]University of Arkansas, [7]University of Texas at Austin,
[8]University of California San Diego, [9]MBZUAI, [10]Oldenburg University.

Abstract

Constructing a robust model that can effectively generalize to test samples under distribution shifts remains a significant challenge in the field of medical imaging. The foundational models for vision and language, pre-trained on extensive sets of natural image and text data, have emerged as a promising approach. It showcases impressive learning abilities across different tasks with the need for only a limited amount of annotated samples. While numerous techniques have focused on developing better fine-tuning strategies to adapt these models for specific domains, we instead examine their robustness to domain shifts in the medical image segmentation task. To this end, we compare the generalization performance to unseen domains of various pre-trained models after being fine-tuned on the same in-distribution dataset and show that foundation-based models enjoy better robustness than other architectures. From here, we further developed a new Bayesian uncertainty estimation for frozen models and used them as an indicator to characterize the model's performance on out-of-distribution (OOD) data, proving particularly beneficial for real-world applications. Our experiments not only reveal the limitations of current indicators like *accuracy on the line* or *agreement on the line* commonly used in natural image applications but also emphasize the promise of the introduced Bayesian uncertainty. Specifically, lower uncertainty predictions usually tend to higher out-of-distribution (OOD) performance.

1 Introduction

Recent years have witnessed tremendous success of foundation models [1–3], which have greatly impacted research in many domains, ranging from understanding language [4], to vision [5]. Such models are pre-trained on massive datasets and have shown encouraging capabilities in performing many tasks, even when only fine-tuned on a small number of samples [6]. Since then, foundation models have presented unprecedented opportunities for researchers to explore more challenging and impactful problems. Among these, medical image understanding [7] has been an attractive topic due to its potential impacts on our society. However, because of the intrinsic differences between medical and natural images, directly applying models pre-trained on natural images to the medical domain may lead to unsatisfactory results [8–10]. Thus, it is imperative to investigate the transferability and robustness of foundation models to unlock their full potential for real-world medical applications.

[*] Co-equal contribution. Corresponding email: Ho_Minh_Duy.Nguyen@dfki.de

R0-FoMo: Workshop on Robustness of Few-shot and Zero-shot Learning in Foundation Models at NeurIPS 2023.

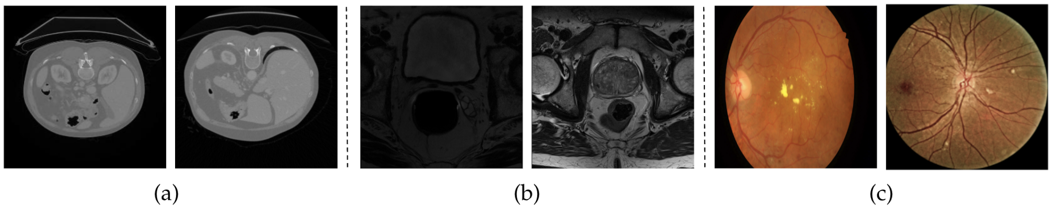

Figure 1: Illustration of domain shift in medical segmentation. From left to right are pairs of liver (a), prostate (b), and diabetic retina images (c) selected from different datasets.

In this work, we investigate the ability to generalize to unseen distributions of various deep learning models, especially large foundation models, in the medical image segmentation task. To this end, we first consider several popular architectures based on UNet and MedSAM. UNet and its derivatives, including UNet++ [11] or TransUNet [12], have conventionally served as prevalent approaches in medical image segmentation. In contrast, MedSAM [13] is a recent method that focuses on *fine-tuning* the Segment Anything Model (SAM) [6]), which is one of the best publicly available models for generic segmentation [14], on a specific medical dataset. We comprehensively evaluate such models on various medical image segmentation tasks by training them on the source domain and then performing evaluations on target domains, which come from different distributions as illustrated in Figure 1. Across all datasets, the experiments showed that fine-tuned foundation models such as MedSAM offer better generalization to unseen domains than traditional models pre-trained on ImageNet. These results demonstrate the promising capabilities of foundation models for real-world deployment.

For a reliable real-world deployment, however, one needs to estimate the model's OOD performance without actually training the model on the target domain. For example, consider a collection of pre-trained models that can segment liver images; given a hospital in a different location, one would like to select a model that will yield the best segmentation results for the patients at this location without training all the models in the collection. However, the lack of labeled samples in the target domain [15, 16] necessitates an effective strategy to model the OOD performance using only unlabeled data. To investigate this problem, we first consider several popular indicators that have shown promising results with natural images, such as the in-domain (ID) performance [15] or the agreement with another network [17]. Interestingly, our findings show that none of such indicators are applicable in the medical image segmentation tasks as they do not hold linear correlations to the OOD performance as expected (Figure 4). This motivates us to introduce a tailored Bayesian uncertainty estimator designed specifically for segmentation tasks, aiming to provide a more dependable indicator for predicting out-of-distribution (OOD) performance. Our experimental results indicate that higher uncertainty in the model's predictions consistently reflects lower OOD performance. In summary, we shed light on the challenges associated with accurately estimating OOD performance in medical image segmentation tasks, underscore the limitations of conventional indicators applied in natural image contexts, and demonstrate the promising results achieved through the proposed Bayesian uncertainty as a surrogate estimator.

2 Methodology

2.1 Notations and Settings

We denote $D = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N} \overset{\text{i.i.d}}{\sim} S$ as a labeled training set of N samples which are independently and identically distributed (i.i.d) sampled from a source domain S. Here, $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ represents a pair consisting of an image (flattened to become a vector) $\boldsymbol{x}_i \in \mathbb{R}^m$ and its label $\boldsymbol{y}_i \in \mathbb{R}^m$. We introduce a deep network denoted as $\boldsymbol{\Phi}(.; \theta) : \mathbb{R}^m \rightarrow \mathbb{R}^m$, parameterized by θ, which maps the images from the set $\boldsymbol{x}_i$ to the desired outputs $\boldsymbol{y}_i$. The primary objective of our learning process is to train a model $\boldsymbol{\Phi}$ using the training dataset D where θ is possibly initialized from foundational models (SAM) or pre-trained models like ImageNet, which are trained on large amounts of natural images. The trained model will then be evaluated for accuracy when applied to samples from a different target domain T. In our setup, the source and target domains, S and T, respectively, are chosen from datasets that pertain to the same organ. However, these datasets experience domain shift issues due to variations in scanner devices or acquisition conditions (Figure 1).

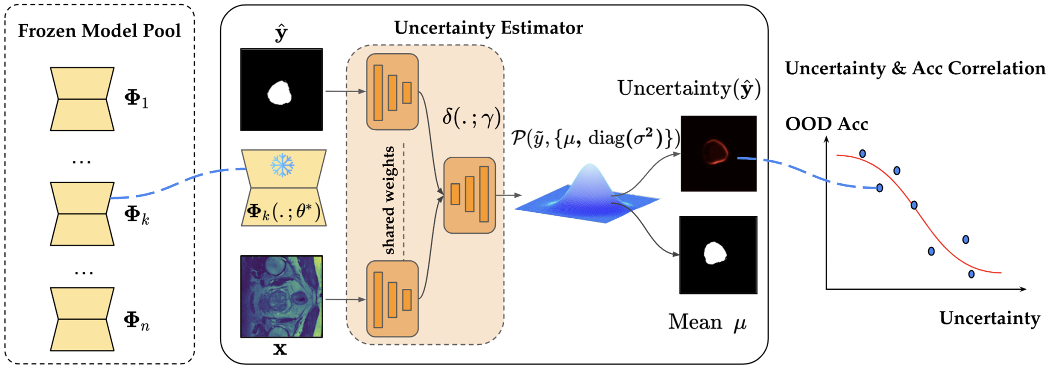

Figure 2: Predicting OOD performance using uncertainties for a pool of frozen models. Each trained network $\boldsymbol{\Phi}_k(.;\theta^*)$ on a source domain is paired with an independent network $\boldsymbol{\delta}(.;\gamma)$ which takes both input images $\boldsymbol{x}$ and output $\hat{\boldsymbol{y}} = \boldsymbol{\Phi}_k(\boldsymbol{x};\theta^*)$ and produces uncertainty for each prediction $\hat{\boldsymbol{y}}$ of $\boldsymbol{\Phi}_k(.;\theta^*)$ in a target domain. The uncertainty values are extracted to estimate a correlation to OOD performance in the target domain for each model.

2.2 Prompt-based Segmentation

We examine the generalization capabilities of SAM for a prompt-based segmentation scenario. We follow the MedSAM method [9] to freeze the image and prompt encoder layers; only mask decoder layers are learned during the fine-tuning phase. The prompt layers use box-based prompts generated by rectangles covering ground-truth regions perturbed with random offsets. We then train it with training data in the single source domain S and take the trained network to predict the test set of the target domain T. Implementation details are presented in Section A.1 Appendix.

2.3 Performance Prediction on Out-of-Distribution under Uncertainty Perspective

Our research proposes a new perspective to explore the relationship between in-distribution (ID) and OOD performance by quantifying the model's uncertainty in the target domains. Unlike previous research [15–17], which mainly focused on image classification tasks and proposed indicators to correlate ID and OOD linearly, our settings in medical image segmentation tasks highlight a crucial difference: none of the previously suggested indicators are applicable in our domain, as demonstrated in Figure 4, where relatively modest Pearson correlation coefficients are observed. This underscores the need for a different approach specific to medical image analysis to capture such complex relationships. Our focus delves into these challenges and provides a proof of concept, indicating that leveraging Bayesian techniques for uncertainty estimation in frozen models can effectively forecast model performance in out-of-distribution scenarios, eliminating the need for labeled samples.

Formulation: Given a trained model $\boldsymbol{\Phi}(,;\theta^*)$, for an input $\boldsymbol{x}_i \in T$, our goal is to model a distribution of $\mathcal{P}(\tilde{\boldsymbol{y}}|\boldsymbol{x}_i)$ rather than a point estimate $\hat{\boldsymbol{y}}_i = \boldsymbol{\Phi}(\boldsymbol{x}_i;\theta^*)$. While Bayesian deep learning-based approaches [18] can provide such uncertainty estimations, they require predefined architectures and have to train models from scratch [19, 20]. However, in our context, accessing complete training data during the pre-trained stage is not feasible, as seen in foundation models, and training costs can be prohibitively high. To tackle this challenge, we turn to the latest post-hoc techniques [18, 21–25] for estimating uncertainty in tasks such as image translation, image enhancement, and depth prediction in self-driving cars. This technique has recently been adapted to active learning research [23] and Vision-Language Models [25]. We tailor these methods for medical segmentation, where regions of interest are often small and surrounded by similar tissue structures. We depict our method in Figure 2.

In particular, we construct an probabilistic model $\boldsymbol{\delta}(.;\gamma)$ to estimate uncertainty for the frozen model $\boldsymbol{\Phi}(.;\theta)$ by producing an independent multivariate Gaussian distribution $\mathcal{P}\left(\tilde{\boldsymbol{y}}; \{\boldsymbol{\mu}_i, \text{diag}(\boldsymbol{\sigma}_i^2)\}\right)$ parameterized by $\{\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2\} = \boldsymbol{\delta}(\{\hat{\boldsymbol{y}}_i, \boldsymbol{x}_i\};\gamma)$ for every input-output pair $\{\boldsymbol{x}_i, \hat{\boldsymbol{y}}_i\}$. This distribution conveys information about the possible values of the reconstructed output $\tilde{\boldsymbol{y}}_i$ and the uncertainty $\boldsymbol{\sigma}_i^2$ of the prediction. We seek to optimize the uncertainty estimator to maximize the data likelihood:

$$
\begin{aligned}
\gamma^* &= \arg\max_{\gamma} \prod_{i=1}^{N} \mathcal{P}\left(\boldsymbol{y}_i; \{\boldsymbol{\mu}_i, \mathrm{diag}(\boldsymbol{\sigma}_i^{-2})\}\right) &(1)\\
&= \arg\max_{\gamma} \prod_{i=1}^{N} \frac{1}{(2\pi)^{m/2}\left(\prod_{j=1}^{m}\sigma_{ij}^2\right)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{y}_i-\boldsymbol{\mu}_i)^\top \mathrm{diag}(\boldsymbol{\sigma}_i^{-2})(\boldsymbol{y}_i-\boldsymbol{\mu}_i)\right) &(2)\\
&= \arg\max_{\gamma} \prod_{i=1}^{N}\prod_{j=1}^{m} \frac{1}{(2\pi)^{1/2}\sigma_{ij}} \exp\left(-\frac{1}{2\sigma_{ij}^2}(\boldsymbol{y}_{ij}-\boldsymbol{\mu}_{ij})^2\right) &(3)\\
&= \arg\min_{\gamma} \sum_{i=1}^{N}\sum_{j=1}^{m} \frac{(\boldsymbol{y}_{ij}-\boldsymbol{\mu}_{ij})^2}{2\sigma_{ij}^2} + \frac{\log\sigma_{ij}^2}{2}. &(4)
\end{aligned}
$$

In order to reconstruct output of $\boldsymbol{\Phi}(,;\theta)$ from $\boldsymbol{\delta}(.;\gamma)$, we can extend Eq.(4) to:

$$
\gamma^* = \arg\min_{\gamma} \sum_{i=1}^{N}\sum_{j=1}^{m} \frac{(\boldsymbol{y}_{ij}-\boldsymbol{\mu}_{ij})^2}{2\sigma_{ij}^2} + \frac{\log\sigma_{ij}^2}{2} + \lambda\|\boldsymbol{\mu}_i - \hat{\boldsymbol{y}}_i\|^2 \tag{5}
$$

where λ indicates the hyper-parameter controlling the trade-off between maximum likelihood and reconstructed output contributions. After optimizing Eq.(5) on samples of source domain S, we then can compute uncertainty for predictions $\hat{\boldsymbol{y}}_t = \boldsymbol{\Phi}(\boldsymbol{x}_t;\theta^*)$ with $\boldsymbol{x}_t \in T$ by:

$$
\mathrm{Uncertainty}(\hat{\boldsymbol{y}}_t) = \boldsymbol{\sigma}_t^2 \in \mathbb{R}^m, \text{ where } \{\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t^2\} = \boldsymbol{\delta}(\{\hat{\boldsymbol{y}}_t, \boldsymbol{x}_t\};\gamma^*). \tag{6}
$$

Generalized Uncertainty with Heavy-Tailed Distribution: Note that Eq. (3) can be seen as a product of univariate Gaussians $\mathcal{N}(\boldsymbol{\mu}_{ij}, \sigma_{ij}^2)$ modeling distributions at the per-pixel level. Therefore, one can extend them to generalized Gaussian distributions [26] $\mathcal{GGD}(\boldsymbol{\mu}_{ij}, \alpha_{ij}, \beta_{ij})$ with scale α_{ij} and shape β_{ij} coefficients used to model heavy-tailed distribution at pixels which usually occur due to the presence of noises or artifacts [22, 24, 25]. The uncertainty for the prediction of $\boldsymbol{x}_t \in T$ in this case is computed as:

$$
\begin{aligned}
&\{\boldsymbol{\mu}_t, \boldsymbol{\beta}_t, \boldsymbol{\alpha}_t\} = \boldsymbol{\delta}(\{\hat{\boldsymbol{y}}_t, \boldsymbol{x}_t\};\gamma^*);\ \boldsymbol{\beta}_t = [\beta_{t1}, ..., \beta_{tm}]\,; \boldsymbol{\alpha}_t = [\alpha_{t1}, ..., \alpha_{tm}] &(7)\\
&\mathrm{Uncertainty}(\hat{\boldsymbol{y}}_t) = \left[\frac{\alpha_{t1}^2\Gamma\left(3\beta_{t1}^{-1}\right)}{\Gamma\left(\beta_{t1}^{-1}\right)}, ..., \frac{\alpha_{km}^2\Gamma\left(3\beta_{tm}^{-1}\right)}{\Gamma\left(\beta_{tm}^{-1}\right)}\right]^\top \in \mathbb{R}^m. &(8)
\end{aligned}
$$

where $\Gamma(z) = \int_0^{\inf} x^{z-1}\exp(-x)dx,\ \forall z > 0$ is the Gamma function.

2.4 Comparison to Existing Models for Post-hoc Uncertainty Quantification

Our formulation is similar to that of existing approaches [22, 24, 25] in that we also estimate the uncertainty of the frozen model $\boldsymbol{\Phi}(.;\theta^*)$ by training an auxiliary network $\boldsymbol{\delta}(.;\gamma)$ with the same objective. However, instead of conditioning only on the output $\hat{\boldsymbol{y}}_i = \boldsymbol{\Phi}(\boldsymbol{x}_i;\theta^*)$, i.e., $\mathrm{Uncertainty}(\hat{\boldsymbol{y}}_i) \sim \boldsymbol{\delta}(\hat{\boldsymbol{y}}_i;\gamma^*)$, we propose a model for uncertainty quantification $\boldsymbol{\delta}(.;\gamma)$ driven by both $\hat{\boldsymbol{y}}_i$ and the original image input $\boldsymbol{x}_i$ for the medical segmentation by $\mathrm{Uncertainty}(\hat{\boldsymbol{y}}_i) \sim \boldsymbol{\delta}(\{\hat{\boldsymbol{y}}_i, \boldsymbol{x}_i\};\gamma^*)$. This is motivated by the following observations.

First, in segmentation settings, the outputs $\hat{\boldsymbol{y}}_i$ produced by the frozen models are simply binary masks and, therefore, there is no surrounding context for $\boldsymbol{\delta}(.;\gamma)$ to learn the maximum likelihood conditions as in Eq.(4). In other words, the model tends to reconstruct only the output of frozen models $\boldsymbol{\Phi}(,;\theta^*)$ by minimizing $\|\boldsymbol{\mu}_i - \hat{\boldsymbol{y}}_i\|^2$ while tending to over-fit the remaining components in Eq.(5). This is essentially different from other methods [22, 25, 24] designed for image enhancement, image translation, or depth estimation, wherein outputs $\hat{\boldsymbol{y}}_i$ are continuous signals and have high mutual information with input images $\boldsymbol{x}_i$.

Second, under a probabilistic view, we have:

$$\mathcal{P}\left(\text{Uncertainty}(\hat{\boldsymbol{y}}), \hat{\boldsymbol{y}} \,|\, \boldsymbol{x}\right) = \mathcal{P}\left(\text{Uncertainty}(\hat{\boldsymbol{y}}) \,|\, \hat{\boldsymbol{y}}, \boldsymbol{x}\right) \cdot \mathcal{P}\left(\hat{\boldsymbol{y}} \,|\, \boldsymbol{x}\right). \tag{9}$$

Both our model $\boldsymbol{\delta}(;\gamma)$ and existing methods estimate $\mathcal{P}\left(\hat{\boldsymbol{y}}|\boldsymbol{x}\right)$ by reconstructing the output of the frozen model $\hat{\boldsymbol{y}} = \boldsymbol{\Phi}(\boldsymbol{x}|\theta^*)$ trained during the pre-training step. However, the remaining factor $\mathcal{P}\left(\text{Uncertainty}(\hat{\boldsymbol{y}}) \,|\, \hat{\boldsymbol{y}}, \boldsymbol{x}\right)$ is learnt by utilizing pairs of $\{\hat{\boldsymbol{y}}, \boldsymbol{x}\}$ in $\boldsymbol{\delta}(;\gamma)$ while other methods approximate $\mathcal{P}\left(\text{Uncertainty}(\hat{\boldsymbol{y}}) \,|\, \hat{\boldsymbol{y}}, \boldsymbol{x}\right) \approx \mathcal{P}\left(\text{Uncertainty}(\hat{\boldsymbol{y}}) \,|\, \hat{\boldsymbol{y}}\right)$. This approximation fails, however, if $\hat{\boldsymbol{y}}$ does not contain sufficient information about the input $\boldsymbol{x}$ as is the case in the segmentation settings.

Finally, instead of computing averaging uncertainty values as prior works, we utilize Otsu's parameter-free thresholding algorithm [27] on the uncertainty matrix. This identifies pixel groups with the highest uncertainty, allowing us to measure their areas. Our approach, particularly effective for small regions of interest in segmentation tasks, exhibits a stronger correlation to OOD scenarios.

Table 1: Performance comparison in domain-shift of prostate segmentation. Results are reported in average 3D IoU score with three training times. The BMC and BIDMC datasets are selected as source domains, and the remaining datasets are used as target domains. Arrows ↑ and ↓ indicate the increase/drop performance between target and source domains.

		UNet (R50)	**Unet ++ (R50)**	**Unet (Eff. Net)**	**UNet ++ (Eff. Net)**	**TransUNet**	**MedSAM**
Same-Dom.	**BMC**	76.77 ± 0.47	75.99 ± 1.21	77.86 ± 1.12	77.63 ± 1.00	65.61 ± 2.72	91.38 ± 0.69
	RUNMC	77.92 ± 1.39	78.62 ± 0.24	79.33 ± 0.92	79.79 ± 0.36	66.17 ± 1.64	83.75 ± 0.87
	BIDMC	71.44 ± 1.28	73.12 ± 0.41	75.55 ± 0.66	74.42 ± 0.45	49.30 ± 8.22	85.04 ± 1.80
	HK	71.34 ± 0.85	70.76 ± 1.48	70.45 ± 1.43	69.69 ± 3.50	44.58 ± 4.49	82.36 ± 1.78
Cross-Dom. $S \rightarrow T$	**BMC → RUNMC**	63.37 ± 1.85 (↓ **13.4**)	62.74 ± 2.67 (↓ **13.3**)	58.29 ± 5.09 (↓ **19.6**)	56.34 ± 5.43 (↓ **21.3**)	21.51 ± 2.28 (↓ **44.1**)	85.23 ± 1.19 (↓ **6.1**)
	BMC → BIDMC	51.22 ± 16.52 (↓ **25.5**)	55.18 ± 3.78 (↓ **20.8**)	46.35 ± 9.48 (↓ **31.5**)	43.21 ± 7.95 (↓ **34.4**)	4.67 ± 0.47 (↓ **60.9**)	84.73 ± 1.05 (↓ **6.6**)
	BMC → HK	56.11 ± 4.93 (↓ **20.7**)	55.34 ± 6.01 (↓ **20.7**)	34.61 ± 6.66 (↓ **43.3**)	36.11 ± 5.29 (↓ **41.5**)	13.11 ± 4.66 (↓ **52.5**)	83.40 ± 0.44 (↓ **8.0**)
	BIDMC → BMC	27.07 ± 1.67 (↓ **44.4**)	24.86 ± 9.09 (↓ **48.3**)	45.33 ± 6.39 (↓ **30.2**)	42.41 ± 7.27 (↓ **32.0**)	6.40 ± 1.33 (↓ **42.9**)	90.77 ± 0.69 (↑ **5.7**)
	BIDMC → RUNMC	4.41 ± 0.69 (↓ **67.0**)	6.37 ± 1.41 (↓ **66.8**)	20.81 ± 9.45 (↓ **54.7**)	18.12 ± 11.04 (↓ **56.3**)	4.19 ± 0.54 (↓ **45.1**)	81.27 ± 1.05 (↓ **3.8**)
	BIDMC → HK	52.83 ± 2.62 (↓ **18.6**)	50.16 ± 2.09 (↓ **23.0**)	47.59 ± 1.52 (↓ **28.0**)	48.23 ± 4.30 (↓ **26.2**)	24.64 ± 4.94 (↓ **24.7**)	81.19 ± 0.92 (↓ **3.9**)

3 Experiment Results

3.1 Datasets and Implementations

We experiment on three segmentation tasks, including Diabetic Retinopathy *(DR) grading-related lesion segmentation* in 2D fundus images, *liver structure segmentation* from 3D CT scans, and *prostate segmentation* from 3D MRI data. Table 5 in the Appendix provides details about each task, including information about the source and target domains, aiming to explore challenges related to domain shifts. We use the default training, testing split in each dataset if available; otherwise, we randomly select 80% total samples for training and 20% remaining for testing. On 3D segmentation problems, we reformulate them as independent 2D slice segmentation and merge results to a single 3D volume to compare with ground truths.

The large vision model SAM [6] is bench-marked by utilizing the `MedSAM` method [13] to fine-tune on a specific medical downstream task. It is important to emphasize that the SAM model was not previously pre-trained with extensive medical images, aligning with the approach taken by pre-trained ImageNet models [28]. We also compare SAM against popular supervised methods such as `TransUNet` with ViT-16 backbone [29], `U-Net` [30] and `U-Net++` [11] with ResNet50 (R50) [31], and Efficient-Net b0 (Eff.Net) [32]. All weights are initialized from ImageNet [28]. Details for uncertainty network $\boldsymbol{\delta}(\{\hat{\mathbf{y}}, \mathbf{x}\}; \gamma^*)$ are presented in Appendix.

3.2 Quantitative Performance on Cross-domain

We report the performance of various model architectures on different medical modalities in Tables 1, 2, and 3. Each model was initially trained and evaluated within the *same domain*. Subsequently, training is conducted on the source domain S, and evaluation is performed on the target domain T, adhering to the *cross-domain* $S \rightarrow T$ setting. Across all scenarios, `MedSAM` consistently demonstrates superior in- and out-of-domain performances, significantly surpassing other models. This noteworthy

Table 2: Performance comparison in domain-shift of DR lesion segmentation. Results are reported in average 2D Dice score with three training times. Arrows ↑ and ↓ indicate the increase/drop performance.

	Same-Domain		**Cross-Domain** $S \rightarrow T$	
	IDRID	**FGADR**	**IDRID → FGADR**	**FGADR → IDRID**
UNet (R50)	35.72 ± 1.35	49.46 ± 1.07	13.51 ± 5.56 (↓ **22.21**)	31.76 ± 1.61 (↓ **17.7**)
Unet ++ (R50)	32.36 ± 3.70	48.73 ± 1.42	10.42 ± 2.19 (↓ **21.9**)	30.60 ± 1.31 (↓ **18.1**)
Unet (Eff.Net)	34.18 ± 2.00	48.78 ± 0.59	12.72 ± 1.24 (↓ **21.5**)	33.61 ± 1.29 (↓ **15.2**)
UNet ++ (Eff.Net)	36.70 ± 1.61	49.88 ± 0.85	20.51 ± 1.41 (↓ **16.2**)	32.49 ± 2.08 (↓ **17.4**)
TransUNet	15.28 ± 1.62	46.59 ± 0.80	11.58 ± 11.04 (↓ **3.7**)	22.52 ± 2.39 (↓ **24.1**)
MedSAM	37.76 ± 1.37	58.49 ± 0.29	44.63 ± 1.42 (↑ 6.9)	39.24 ± 0.62 (↓ **19.25**)

Table 3: Performance comparison in domain-shift of liver segmentation. Results are computed by average 2D Dice score in three training times. Arrows ↑ and ↓ indicate the increase/drop performance.

	Same-Domain		**Cross-Domain** $S \rightarrow T$	
	FLARE	**LiTS**	**FLARE → LiTS**	**LiTS → FLARE**
UNet (R50)	94.35 ± 1.16	95.69 ± 0.09	82.28 ± 4.41 (↓ **12.1**)	95.57 ± 0.39 (↓ **0.1**)
Unet ++ (R50)	96.08 ± 0.40	95.84 ± 0.12	72.51 ± 5.97 (↓ **23.6**)	95.41 ± 0.45 (↓ **0.4**)
Unet (Eff. Net)	95.11 ± 0.18	95.86 ± 0.23	68.91 ± 7.06 (↓ **26.2**)	95.04 ± 0.63 (↓ **0.8**)
UNet ++ (Eff. Net)	95.23 ± 0.98	95.57 ± 0.40	67.32 ± 3.27 (↓ **27.91**)	95.0 ± 0.95 (↓ **0.6**)
TransUNet	92.01 ± 0.78	92.22 ± 3.22	61.69 ± 1.17 (↓ **30.3**)	93.2 ± 1.68 (↑ **1.0**)
MedSAM	92.47 ± 0.02	97.80 ± 0.01	97.53 ± 0.03 (↑ 5.1)	92.17 ± 0.24 (↓ **5.6**)

achievement highlights MedSAM's robust OOD generalization capabilities, proven effective across both balanced (prostate modality) and imbalanced (diabetic retinopathy lesion and liver modalities) datasets. Further detailed analysis is available in the Appendix. To our latest knowledge, we first examine the robustness of the SAM model under domain shift in medical segmentation tasks, given models are fine-tuned on medical source domains [33, 34].

3.3 Estimating Out-of-Distribution Performance via Uncertainty Perspective

We now investigate different indicators to characterize the models' OOD properties. We consider both popular indicators used in the natural image domains, such as the ID performance [15], the agreement [17], and our proposed Bayesian uncertainty.

Figure 3 illustrates a strong correlation between uncertainty and OOD performance in three settings of prostate segmentation (Table 1), underscoring that elevated uncertainty values tend to align with diminished OOD performance. Additionally, when comparing different models on the same dataset, the OOD performance gaps between these models are found to be correlated with the gaps in their uncertainties. However, it is important to highlight that uncertainty only approximated the true error in this context. As a result, even when two models exhibit similar OOD performances, there may still be some subtle differences in their uncertainties, reflecting nuanced variations in their predictive capabilities. This phenomenon presents an interesting future research direction for accurately estimating a model OOD performance.

Comparing with Other Correlations: Figure 4 reports the correlations between the OOD performance and the ID performance or the ID agreement when using BMC as the source domain and BIDMC as the target domain. By varying the training configurations, such as learning rates, epochs, etc., we plot the OOD performance against the ID indicator for each model. Afterward, we compute Pearson correlation coefficients between the OOD performance and ID indicators. This analysis aims to substantiate whether linear correlations, as proposed in previous studies [15, 17], are indeed observed. Nevertheless, none of these indicators proves effective in adequately characterizing the OOD performance, as they only yield relatively low Pearson correlation coefficients. Consequently, we can infer that the currently employed natural image indicators are not well-suited for our specific medical image segmentation task.

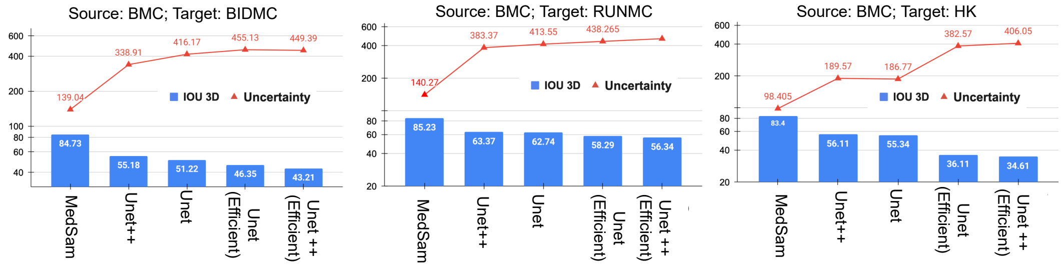

Figure 3: Visualization of uncertainty and OOD performance on three cross-domain experiments in prostate segmentation. Uncertainty exhibits a strong correlation with OOD performance, indicating that higher uncertainty values tend to correspond to lower OOD performance.

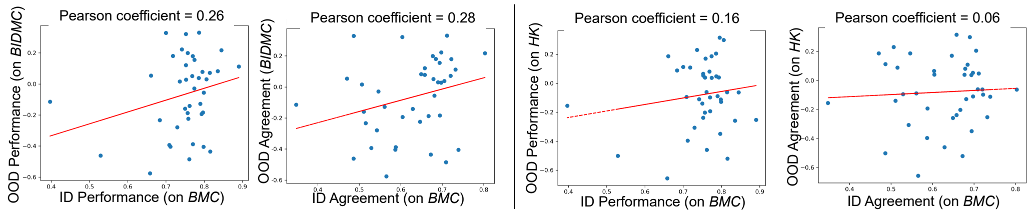

Figure 4: OOD performance/agreement versus ID performance/agreement on BMC and HK datasets (left) and on BMC and BIDMC datasets (right).

Comparing with Uncertainty Baselines: We furthermore compare our approach against two different baseline groups, including Bayesian identity mapping [22] and perturbing the original images to gauge the uncertain level for the SAM model. The former group, known as BayesCap [22], calibrates the uncertainties of a pre-trained model without the need to train itself on a large-scale dataset. The latter is based on *test-time modification* techniques [35–37]. Herein, we employ two types of modifications: test-time data augmentation (TTDA) [35, 36] and dropout [38] before the final prediction [39, 40]. Two classes of data augmentation are adapted for baselines, including pixel-wise noise (TTDAp) and color jittering (TTDAc). Table 4 illustrates various methods' performance in three cross-domain prostate segmentation settings. The results are presented as mean square errors between the uncertainty values, estimated using the Otsu thresholding algorithm, and the actual errors of the model. Notably, our approach consistently outperforms others, attaching the top records across all three Out-of-Domain (OOD) settings.

4 Conclusion

In this study, we first conducted a systematic investigation into the performance of the foundation model SAM under domain-shift segmentation tasks. Our findings reveal that this foundation model demonstrates better generalization capabilities than other methods initialized from pre-trained ImageNet. Additionally, we show the correlation between model uncertainties and their out-of-distribution performance; thereby, lower uncertainty predictions tend to reflect a higher performance in OOD. To achieve this, we constructed a post-hoc estimation approach tailored for pre-trained deterministic models. These models have demonstrated the ability to consistently generate well-calibrated uncertainty estimates across various segmentation scenarios, proving their utility in aiding experts during the decision-making process. One limitation of our algorithm is that the uncertainty can only approximate the true errors; therefore, it may not be suitable to compare models whose performances have small margins. In future work, we plan to improve this issue as well as explore the proposed method in a broader range of foundation methods with diverse settings to gain better insights into the behaviors of the proposed algorithms.

Table 4: Comparison among uncertainty approaches measured by mean squared error. Smaller is better.

Method	BMC → RUNMC	BMC → BIDMC	BMC → HK
Our	**0.012**	**0.011**	**0.018**
BayesCap	0.015	0.017	0.021
TTDAp	0.247	0.326	0.387
TTDAc	0.212	0.248	0.291
DropOut	0.197	0.227	0.271

Acknowledgements

This research has been supported by the pAItient project (BMG, 2520DAT0P2), Ophthalmo-AI project (BMBF, 16SV8639), and Endowed Chair of Applied Artificial Intelligence, Oldenburg University, Germany. Tan Ngoc Pham would like to thank the Vingroup Innovation Foundation (VINIF) for the Master's training scholarship program. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Duy M. H. Nguyen and Vinh Tong. Mathias Niepert acknowledges funding by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC and support by the Stuttgart Center for Simulation Science (SimTech).

References

[1] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[2] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[5] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[7] Martin J Willemink, Holger R Roth, and Veit Sandfort. Toward foundational deep learning models for medical imaging in the new era of transformer networks. *Radiology: Artificial Intelligence*, 4(6):e210284, 2022.

[8] Duy MH Nguyen, Thu T Nguyen, Huong Vu, Quang Pham, Manh-Duy Nguyen, Binh T Nguyen, and Daniel Sonntag. Tatl: task agnostic transfer learning for skin attributes detection. *Medical Image Analysis*, 78:102359, 2022.

[9] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023.

[10] Peilun Shi, Jianing Qiu, Sai Mu Dalike Abaxi, Hao Wei, Frank P-W Lo, and Wu Yuan. Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation. *arXiv preprint arXiv:2304.12637*, 2023.

[11] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018.

[12] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[13] Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, et al. Segment anything model for medical images? *arXiv preprint arXiv:2304.14660*, 2023.

[14] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021.

[15] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pages 7721–7735. PMLR, 2021.

[16] Weijian Deng, Stephen Gould, and Liang Zheng. On the strong correlation between model invariance and generalization. *Advances in Neural Information Processing Systems*, 35:28052–28067, 2022.

[17] Christina Baek, Yiding Jiang, Aditi Raghunathan, and J Zico Kolter. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. *Advances in Neural Information Processing Systems*, 35:19274–19289, 2022.

[18] Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. *Advances in Neural Information Processing Systems*, 34:20089–20103, 2021.

[19] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 270–279, 2017.

[20] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[21] Runa Eschenhagen, Erik Daxberger, Philipp Hennig, and Agustinus Kristiadi. Mixtures of laplace approximations for improved post-hoc uncertainty in deep learning. *arXiv preprint arXiv:2111.03577*, 2021.

[22] Uddeshya Upadhyay, Shyamgopal Karthik, Yanbei Chen, Massimiliano Mancini, and Zeynep Akata. Bayescap: Bayesian identity cap for calibrated uncertainty in frozen neural networks. In *European Conference on Computer Vision*, pages 299–317. Springer, 2022.

[23] Vikrant Rangnekar, Uddeshya Upadhyay, Zeynep Akata, and Biplab Banerjee. Usim-dal: Uncertainty-aware statistical image modeling-based dense active learning for super-resolution. *arXiv preprint arXiv:2305.17520*, 2023.

[24] Uddeshya Upadhyay, Sairam Bade, Arjun Puranik, Shahir Asfahan, Melwin Babu, Francisco Lopez-Jimenez, Samuel Asirvatham, Ashim Prasad, Ajit Rajasekharan, Samir Awasthi, et al. Hypuc: Hyperfine uncertainty calibration with gradient-boosted corrections for reliable regression on imbalanced electrocardiograms. *Transactions on Machine Learning Research*, 2023.

[25] Uddeshya Upadhyay, Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Probvlm: Probabilistic adapter for frozen vison-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1899–1910, 2023.

[26] Alex Dytso, Ronit Bustin, H Vincent Poor, and Shlomo Shamai. Analytical properties of generalized gaussian distributions. *Journal of Statistical Distributions and Applications*, 5(1): 1–40, 2018.

[27] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[28] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[29] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representatio*, 2021.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[32] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[33] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023.

[34] Yu Qiao, Chaoning Zhang, Taegoo Kang, Donghun Kim, Shehbaz Tariq, Chenshuang Zhang, and Choong Seon Hong. Robustness of sam: Segment anything under corruptions and beyond. *arXiv preprint arXiv:2306.07713*, 2023.

[35] Murat Seckin Ayhan and Philipp Berens. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *Medical Imaging with Deep Learning*, 2022.

[36] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sebastien Ourselin, and Tom Vercauteren. Test-time augmentation with uncertainty estimation for deep learning-based medical image segmentation. 2022.

[37] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

[38] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[39] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.

[40] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[42] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[43] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[44] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020.

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[46] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

[47] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[48] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[49] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[50] Yi Zhou, Boyang Wang, Lei Huang, Shanshan Cui, and Ling Shao. A benchmark for studying diabetic retinopathy: segmentation, grading, and transferability. *IEEE Transactions on Medical Imaging*, 40(3):818–828, 2020.

[51] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018.

[52] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

[53] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical Image Analysis*, 84:102680, 2023.

[54] Quande Liu, Qi Dou, Lequan Yu, and Pheng Ann Heng. Ms-net: multi-site network for improving prostate segmentation with heterogeneous mri data. *IEEE transactions on medical imaging*, 39(9):2713–2724, 2020.

Supplementary Material for “On the Out of Distribution Robustness of Foundation Models in Medical Image Segmentation"

A Background

A.1 SAM Architecture

SAM [6] was introduced by Meta AI to improve segmentation performance across a wide range of images. SAM utilizes a transformer-based architecture [41], which has shown impressive achievement in both natural language processing [42] and computer vision [29]. In general, SAM is a Vision Transformer (ViT)-based model, which consists of three components: an image encoder, a prompt encoder, and a mask decoder. The image encoder is based on ViT [29], which is pre-trained with masked auto-encoder (MAE) [43]. Two sets of prompts can be considered, including sparse prompts (points, boxes, text) and dense prompts (masks). To represent the first two prompts, points, and boxes, SAM employs positional encoding [44] combined with learned embeddings. The text prompt is encoded by the pre-trained text-encoder CLIP [45]. The mask prompt has the same spatial resolution as the input image, and it is encoded using convolutions and summed element-wise with the image embedding. The mask decoder uses a lightweight network consisting of two transformer layers and a dynamic mask prediction head with an Intersection-over-Union (IoU) score regression head. SAM is trained in a supervised learning manner on a large-scale SA-1B dataset with over 1 billion masks from 11 million natural images using a linear combination of Dice loss [46] and Focal loss [47].

A.2 Implementation Details

We use `MedSAM` to fine-tune the SAM model on a specific medical downstream task. The Adam optimizer is utilized for training networks with a combination of Dice loss and Cross-entropy. The learning rates for problems are selected from a set of $\{1e{-}4, 3e{-}4, 5e{-}4, 1e{-}5\}$ depend on validation performance.

To gauge the extent of uncertainty arising from the foundational models, we utilize a deep architecture named $\boldsymbol{\delta}(\{\hat{\mathbf{y}}, \mathbf{x}\}; \gamma^*)$, inspired by the ResNet family [31] with 18 convolutional layers. In our experiment, we use Eqs. (7),(8) for training with $\boldsymbol{\beta}_t$ fixed as $\mathbf{2_m}$. The initial input for the model $\boldsymbol{\delta}$ is a joint feature vector constructed by concatenating the original image $\mathbf{x}$ with the prediction mask $\hat{\mathbf{y}}$ after they have passed through two different convolutional layers. Additionally, the output of the model $\boldsymbol{\delta}$ is also combined with the aforementioned joint feature vector. These combined features then undergo processing through distinct convolutional operators, ultimately producing parameters representing uncertainty distributions as described in Eq.(6). In the end, the uncertainty model $\boldsymbol{\delta}$ is trained with Adam optimizer [48] with the learning rate of $1e^{-4}$ in 50 epochs and cosine annealing strategy [49] for the learning rate warm-up.

B Datasets

Table 5 overviews used dataset in our experiment along with the modality and image types. Each task has at least two datasets, one for a source domain and the remaining for a target domain.

Table 5: Overview datasets used in our experiment.

Task	Objects	Datasets	Modality	# Images
DR Lesion Segmentation	HE, SE, EX, MA	FGADR [50]	2D Fundus	1842
		IDRiD [51]	2D Fundus	81
Liver Segmentation	Liver	FLARE [52]	3D CT	50
		LiTS [53]	3D CT	130
Prostate Segmentation	Prostate	BMC [54]	3D MRI	30
		BIDMC [54]	3D MRI	12
		RUNMC [54]	3D MRI	30
		HK [54]	3D MRI	12

C Experiment Details & Analyses

C.1 Further Analysis on Cross-domain Performance

The performance comparisons are presented in Tables 1, 2, and 3 for three different medical modalities: Prostate, DR, and Liver, respectively. For each dataset, we report the performance on both Unet-based models (including Unet (R50), Unet++ (R50), Unet (Eff.Net), Unet++ (Eff.Net)), and SAM-based model (i.e., MedSAM). Each model was initially trained and tested within the same domain. It is then trained on source domain S and tested on target domain T under the cross-domain $S \rightarrow T$ settings.

Table 1 demonstrates that MedSAM not only outperforms other models but also exhibits superior generalization with minimal performance gaps between the same domain and cross-domain.

Four datasets (BMC, BIDMC, RUNMC, and HK) on Prostate modality are well-balanced in terms of the number of images. Thus, the domain shift has been clearly observed, as shown in Table 1. On the other hand, the DR Lesion datasets (FGADR and IDRiD) and Liver datasets (FLARE and LiTS) suffer from heavy class imbalance, as indicated in Table 5. Notably, the Liver datasets consist of 3D CT Scans processed slide-by-slide, with FLARE having 3,080 slides and LiTS having 25,660 slides. In Table 2, MedSAM achieves state-of-the-art performance with small domain gaps. Due to the large imbalance between LiTS (25,660 slides) and FLARE (3,080 slides), models trained on LiTS, whether Unet-based or SAM-based, demonstrate better generalization compared to those trained on FLARE, as demonstrated in Table 3, except MedSAM. While MedSAM shows a great improvement on small datasets such as FLARE (i.e., FLARE $\rightarrow$ LiTS), its generalization is dropped on large dataset LiTS.

C.2 Qualitative Results

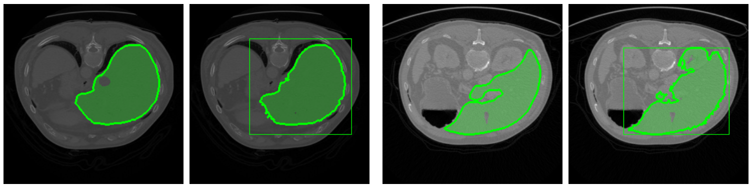

Figure 5: A visual demonstration on MedSAM's performance in the same- and cross-domain. The two on the left depict the true mask and MedSAM's prediction mask in the same domain (FLARE). The other two display the true mask and MedSAM's prediction mask in an out-of-domain scenario (FLARE $\rightarrow$ LiTs). Best views in color with zoom.

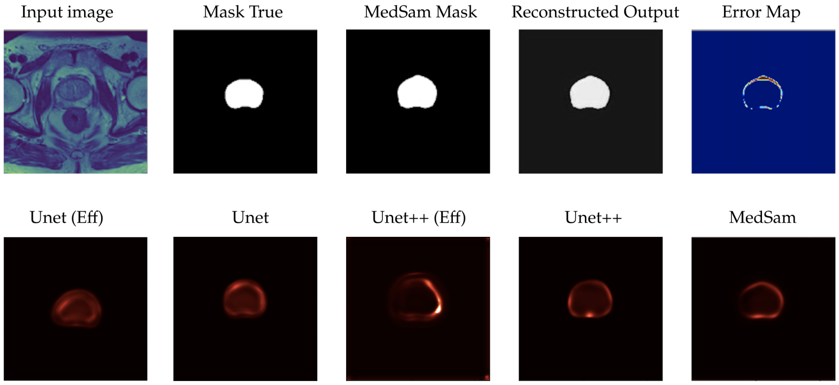

Figure 6: A qualitative comparison among various foundation models. Respectively, from left to right, the top row demonstrates the original image, the true mask for that image, MedSAM's prediction mask, the reconstructed mask from the uncertainty model, and the error map between the true mask and MedSAM's prediction mask. The bottom row ranks the uncertain levels of different foundation models from the most uncertainties (leftmost) to the fewest uncertainties (rightmost), whereas the fewer, the better. Following this, the prediction mask derived from MedSAM has the fewest uncertainties.

In Figure 5, we present a selection of predictions generated by the MedSAM model in both same and cross-domain settings, showcasing the segmentation results for lung, prostate, and soft exudates. Given the prompt-based nature of MedSam, we simulate bounding boxes-based prompts as [9]. Our observations indicate that MedSAM is capable of producing satisfactory masks for prostate structures and DR lesions (shown in the second and bottom rows) despite the challenging conditions of small object sizes and domain shifts. However, when applied to the LiTS dataset, MedSAM tends to exhibit tendencies of over-segmenting boundaries or missing structures at the bends of objects.

We offer illustrative examples of uncertainty estimations generated by our algorithm, showcased in Figure 6, across various trained models. It can be seen the MedSam model has the highest relevant uncertainty regions compared to the Error map among models.