Tyche: Stochastic In-Context Learning for Medical Image Segmentation

Marianne Rakic
MIT, CSAIL & MGH
mrakic@mit.edu

Hallee E. Wong
MIT, CSAIL & MGH

Jose Javier Gonzalez Ortiz
MosaicML DataBricks
& MIT, CSAIL

Beth A. Cimini
Broad Institute
of MIT and Harvard

John V. Guttag
MIT, CSAIL

Adrian V. Dalca
MIT, CSAIL
& MGH, HMS

Abstract

Existing learning-based solutions to medical image segmentation have two important shortcomings. First, for most new segmentation tasks, a new model has to be trained or fine-tuned. This requires extensive resources and machine-learning expertise, and is therefore often infeasible for medical researchers and clinicians. Second, most existing segmentation methods produce a single deterministic segmentation mask for a given image. In practice however, there is often considerable uncertainty about what constitutes the correct segmentation, and different expert annotators will often segment the same image differently. We tackle both of these problems with Tyche*, a framework that uses a context set to generate stochastic predictions for previously unseen tasks without the need to retrain. Tyche differs from other in-context segmentation methods in two important ways. (1) We introduce a novel convolution block architecture that enables interactions among predictions. (2) We introduce in-context test-time augmentation, a new mechanism to provide prediction stochasticity. When combined with appropriate model design and loss functions, Tyche can predict a set of plausible diverse segmentation candidates for new or unseen medical images and segmentation tasks without the need to retrain. The Tyche code is available at:* https://tyche.csail.mit.edu/.

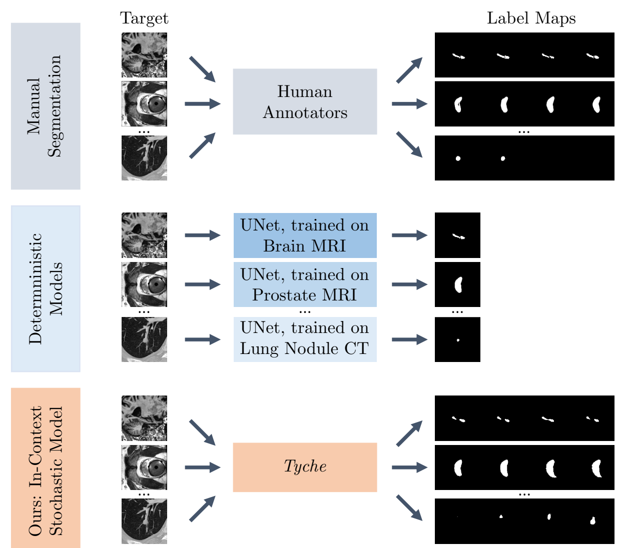

Figure 1. **Tyche: the first in-context stochastic segmentation framework.** Human annotators (top) can handle a wide variety of tasks, and different annotators often produce differing segmentations. Existing automated methods (middle) are typically task-specific and provide only one segmentation per image. *Tyche* (bottom) can capture the disagreement among annotators across many modalities and anatomies without retraining or fine-tuning.

1. Introduction

Segmentation is a core step in medical image analysis, for both research and clinical applications. However, current approaches to medical image segmentation fall short in two key areas. First, segmentation typically involves training a new model for each new modality and biomedical domain, which quickly becomes infeasible given the resources and expertise available in biomedical research and clinical environments. Second, models most often provide a single solution, whereas in many cases, the target image contains ambiguous regions, and there isn't a *single* correct segmentation. This ambiguity can arise from noisy or low contrast images, variation in the task definition, or human raters' interpretations and downstream goals [13, 57]. Failure to take this ambiguity into account can affect downstream analysis, diagnosis, and treatment.

Recent work tackles these issues separately. *In-context learning* (ICL) methods generalize to unseen medical image segmentation tasks, employing an input *context* or *prompt* to guide inference [20, 130, 131]. These methods are deterministic and predict a *single* segmentation for a given input image and task.

Separately, stochastic or probabilistic segmentation methods output multiple plausible segmentations at inference, reflecting the task uncertainty [12, 69, 97]. Each such model is trained for a specific task, and can only output multiple plausible segmentations at inference for that task. Training or fine-tuning a model for a new task requires technical expertise and computational resources that are often unavailable in biomedical settings.

We present *Tyche*, a framework for stochastic ICL medical image segmentation (Figure 1). *Tyche* includes two variants for different settings. The first, *Tyche-TS* (Train-time Stochasticity), is a system explicitly designed to produce multiple candidate segmentations. The second, *Tyche-IS* (Inference-time Stochasticity), is a test-time solution that leverages a pretrained deterministic ICL model.

Tyche takes as input the image to be segmented (target), and a *context set* of image-segmentation pairs that defines the task. This enables the model to perform unseen segmentation tasks upon deployment, omitting the need to train new models. *Tyche-TS* learns a *distribution of possible label maps*, and predicts a set of plausible stochastic segmentations. *Tyche-TS* encourages *diverse* predictions by enabling the internal representations of the different predictions to interact with each other through a novel convolutional mechanism, a carefully chosen loss function and noise as an additional input. In *Tyche-IS*, we show that applying test-time augmentation to both the target and context set in combination with a trained ICL model leads to competitive segmentation candidates.

We make the following contributions.

- We present the first solution for probabilistic segmentation for ICL. We develop two variants to our framework: *Tyche-TS* that is trained to maximize the quality of the best prediction, and *Tyche-IS*, that can be used straightaway with an existing ICL model.
- For *Tyche-TS*, we introduce a new mechanism, *SetBlock*, to encourage diverse segmentation candidates. *Tyche-TS* is simpler than existing stochastic methods, predicting all the segmentation candidates in a single forward pass.
- Through rigorous experiments and ablations on a set of twenty unseen medical imaging tasks, we show that both variants of *Tyche* produce solutions that outperform existing in-context and interactive segmentation benchmarks, and can match the performance of specialized stochastic networks trained on specific datasets.

2. Related Work

Biomedical segmentation is a widely-studied problem, with recent methods dominated by UNet-like architectures [7, 53, 109]. These models tackle a wide variety of tasks, such as different anatomical regions, different structures to segment within a region, different image modalities, and different image settings. With most methods, a new model has to be trained or fine-tuned for each combination of these. Additionally, most models don't take into account image ambiguity, and provide a single deterministic output.

Multiple Predictions. Uncertainty estimation can help users decide how much faith to put in a segmentation [27] and guide downstream tasks. Uncertainty is often categorized into aleatoric, uncertainty in the data, and epistemic, uncertainty in the model [30, 63]. In this work, we focus on aleatoric uncertainty. Medical images are also heteroscedastic in that the degree of uncertainty varies across the image.

Different strategies exist to capture uncertainty. One can assign a probability to each pixel [47, 54, 64, 78], or use contour strategies and difference loss functions to predict the largest and smallest plausible segmentations [75, 135]. These strategies however do not capture the correlations across pixels. To address this, some methods generate multiple plausible label maps given an image [12, 69, 70, 97, 133]. To achieve this, one can directly model pixel correlations, such as through a multivariate Gaussian distribution (with low rank) covariance [97], or more complex distributions [16]. Alternatively, various frameworks combine potentially hierarchical representations for UNet-like architectures with variational auto-encoders [12, 69, 70]. More recently, diffusion models have been used for ensembling [133] or to produce stochastic segmentations [107, 136]. Some methods explicitly model the different annotators to capture ambiguity [50, 102, 113, 126]. But these methods do not apply to our framework where the number of annotators and their characteristics are unknown.

Most of the models above involve sophisticated modeling or lengthy runtimes, and need to be trained on each segmentation task. In *Tyche*, we build on intuition across these methods, but combine a more efficient mechanism with an ICL strategy to predict segmentation candidates.

In-context Learning. Few-Shot frameworks use a small set of examples to generalize to new tasks [32, 82, 100, 103, 115, 119, 137], sometimes by fine-tuning an existing pretrained model [33, 101, 122, 127]. In-context learning segmentation methods (ICL) use a small set of examples directly as input to infer label maps for a task [10, 20, 65, 131]. This enables them to generalize to new tasks. For ex-

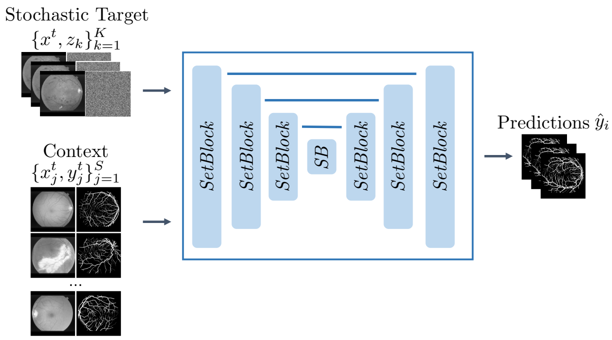

Figure 2. **Tyche Model Schematic.** The target x^t, context set $(x_j^t, y_j^t)_{j=1}^S$, and noise images $\{z_k\}_{k=1}^K$ are inputs to the network. The architecture employs UNet-like levels, but uses *SetBlocks* that enable interactions between the context set and the target segmentation candidates.

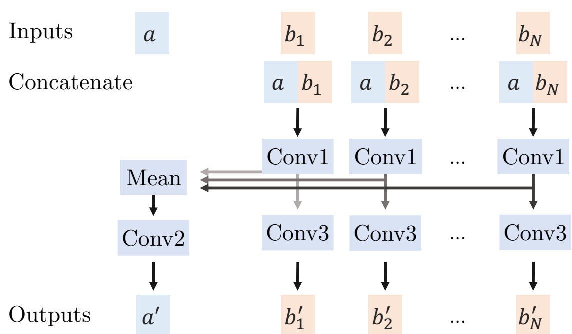

Figure 3. **CrossBlock Mechanism** The CrossBlock involves interactions between a single feature and a set of features and outputs new feature for the target and new features for each.

ample, UniverSeg uses an enhanced UNet-based architecture to generalize to medical image segmentation tasks unseen during training [20]. We build on these ideas to enable segmentation of new tasks without the need to re-train, but expand this paradigm to model stochastic segmentations.

Test Time Augmentation. The test-time augmentation (TTA) strategy uses perturbations of a test input and ensembles the resulting predictions. Existing TTA frameworks model accuracy [35, 66, 118, 121], robustness [26], and estimates of uncertainty [6, 92]. Test-time augmentation has been applied to diverse anatomies and modalities including brain MRI and retinal fundus [4, 6, 51, 55, 99, 129]. Prior work has formalized the variance of a model's predictions over a set of input transformations as capturing aleatoric uncertainty [6, 128, 129].

Tyche's use of TTA is distinct from prior work. Instead of ensembling segmentations over perturbations of a test input or pixel-wise estimates of uncertainty, *Tyche* extends TTA to the ICL setting and uses the individual TTA predictions to model uncertainty.

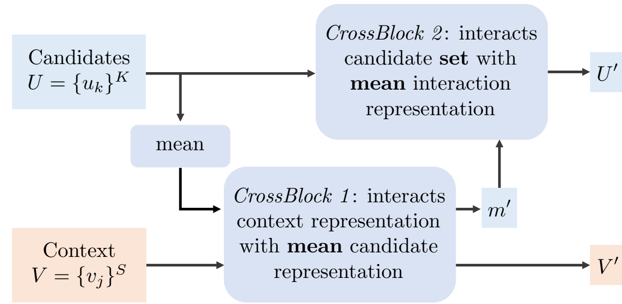

Figure 4. **SetBlock Mechanism**. *SetBlock* enables interactions between the **set** of features from the context set and the **set** of features from the prediction candidates. It outputs two sets of features, one for the context and one for the prediction candidates.

3. Method

For segmentation task t, let $\{(x_j^t, y_j^t)\}_{j=1}^N$ be a dataset with images x^t and label maps y^t. Typical segmentation models learn a different function $\hat{y}^t = g_{\theta^t}(x^t)$ with parameters θ^t for each task t, where $\hat{y}^t$ is a single segmentation map prediction.

We design *Tyche* as an in-context learning (ICL) model using a *single* function for all tasks:

$$\hat{y}_k^t = f_\theta(x^t, z_k, \mathcal{S}^t). \tag{1}$$

This function, with *global* parameters θ, captures a *distribution of label maps* $\{\hat{y}_k^t\}_{k=1}^K$, given target x^t, context set $\mathcal{S}^t = \{x_j^t, y_j^t\}_{j=1}^S$ defining task *t*, and noise $z_k \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$. We use this modelling strategy in two ways: we either explicitly train a network to approximate the model $f_\theta(\cdot)$ in *Tyche-TS*, or design a test-time strategy to approximate $f_\theta(\cdot)$ using an existing (pretrained) deterministic ICL network in *Tyche-IS*.

3.1. Tyche-TS

In *Tyche-TS*, we explicitly train a neural network for $f_\theta(\cdot)$ that can make different predictions given the same image input x^t but different noise channels z_k. We model interaction between predictions, and employ a loss that encourages diverse solutions (Figure 2).

3.1.1 Neural Network

We use a convolutional architecture focused on interacting representations of sets of flexible sizes using a modified version of the usual UNet structure [109].

Inputs. *Tyche-TS* takes as input the target x^t, a set of K Gaussian noise channels z_k, and a context set, $\mathcal{S}^t$.

Layers. Each level of the UNet takes as input a set of K candidate representations and S context representations. We design each level to encourage communication between the intermediate elements of the sets, and between the two features of the segmentation candidates. The size K is flexible and can vary with iterations.

SetBlock. We introduce a new operation called *SetBlock*, which interacts the candidate representations $U = \{u_i\}_{i=1}^K$, with the context representations $V = \{v_i\}_{i=1}^S$, illustrated in Figure 4. We use the *CrossBlock* [20] as a building block for this new layer. The CrossBlock$(u, V) \rightarrow (u', V')$ compares an image representation u to a context set representation V through convolutional and averaging operations, and outputs a new image representation u' and a new set representation V' (Figure 3). SetBlock$(U, V) \rightarrow (U', V')$ builds on CrossBlock and performs a set to set interaction of the entries of U and V:

$$\bar{u} = 1/m \textstyle\sum_{i=1}^m u_i \tag{2}$$

$$\bar{u}', V' = \text{CrossBlock}(\bar{u}, V) \tag{3}$$

$$u_i' = \text{Conv}_m\left(u_i || \bar{u}'\right), \quad i = 1, \dots, K \tag{4}$$

$$u_i' = \text{Conv}_u\left(u_i'\right), \quad i = 1, \dots, K \tag{5}$$

$$v_i' = \text{Conv}_v\left(v_i\right), \quad i = 1, \dots, S, \tag{6}$$

where $||$ is the concatenation operation along the feature dimension. The CrossBlock interacts the context representation with the **mean** candidate. The $Conv_m$ step communicates this result to all candidate representation. $Conv_u$ and $Conv_v$ then update all representations. All convolution operations include a non-linear activation function.

3.1.2 Best candidate Loss

Typical loss function compute the loss of a single prediction relative to a single target, but *Tyche-TS* produces multiple predictions and has one or more corresponding label maps.We optimize

$$\mathcal{L}(\theta; \mathcal{T}) = \mathbb{E}_{t \in \mathcal{T}}\left[\mathbb{E}_{(x^t, y_r^t), \mathcal{S}^t}\left[\mathcal{L}_{seg}\left(\{\hat{y}_k\}, y_r^t\right)\right]\right], \tag{7}$$

with

$$\mathcal{L}_{seg}(\{\hat{y}_k\}, y_r^t) = \min_k \; \mathcal{L}_{Dice}\left(y_k, y_r^t\right), \tag{8}$$

where y_r^t is a segmentation from rater r, and $\mathcal{L}_{Dice}$ is a weighted sum of soft Dice loss [96] and binary cross-entropy. By only back-propagating through the best prediction among K candidates, the network is encouraged to produce diverse solutions [23, 43, 68, 83].

3.1.3 Training Data

We employ a large dataset of single- and multi-rater segmentations across diverse biomedical domains. We then use data augmentation [20], as described in B.2.

We add synthetic multi-annotator data by modelling an image as the average of four blobs representing four raters (Figure 8). Each blob is white disk b_i deformed by a random smoothed deformation field ϕ_i. The synthetic image is a noisy weighted sum of raters: $\sum_{i=1}^4 w_i(b_i \circ \phi_i)$ where $\circ$ represents the spacial warp operation.

3.1.4 Implementation Details

We use a UNet-like architecture of 4 *SetBlock* layers for the encoder and decoder, with 64 features each and Leaky ReLU as activation function. We use the Adam optimizer and a learning rate of 0.0001. At training, we have a fixed number of candidates per sample $K_{\text{tr}} = 8$. At inference, we consider different numbers of candidates.

3.2. Tyche-IS

In *Tyche-IS*, we first train (or use an existing trained) *deterministic* ICL segmentation system $\hat{y}^t = h_\theta(x^t, \mathcal{S}^t)$. We then introduce a *test-time* in-context augmentation strategy to provide stochastic predictions:

$$\hat{y}_k^t = f_\theta(x^t, z_k, \mathcal{S}^t) \tag{9}$$

$$= h_\theta(aug(x^t, z_k, \mathcal{S}^t)), \tag{10}$$

with $\tilde{x}^t, \tilde{\mathcal{S}}^t = aug(x^t, z_k, \mathcal{S}^t)$, an augmentation function.

3.2.1 Augmentation Strategy

Test time augmentation for single task networks $y = g_t(x)$ applies different transforms to an input image x:

$$\tilde{x}_k = a_\phi(x, z_k), \tag{11}$$

where ϕ are augmentation parameters and z_k is a random vector. A final prediction is then obtained by combining several predictions of augmented images. Most commonly, the combining function averages the predictions:

$$y = \frac{1}{k}\sum_k g_t(\tilde{x}_k), \tag{12}$$

where the sum operates pixel-wise.

We introduce in-context test-time augmentation (IC-TTA) as another mechanism to generate diverse stochastic predictions.

We apply augmentation to both the test target x^t and the context set $\mathcal{S}^t$:

$$(\tilde{x_i}^t, y_i^t) = (a_\phi(x_i^t), y^t) \tag{13}$$

$$\tilde{\mathcal{S}}^t = \{a_\phi(x_j^t), y_j^t\}_{j=1}^S. \tag{14}$$

We repeat this process K_i times to obtain K_i stochastic predictions:

$$\hat{y}_k = f_\theta(\tilde{x_i}^t, z_k, \tilde{\mathcal{S}}^t) \tag{15}$$

We only apply intensity based transforms, to avoid the need to invert the predicted segmentations back. We apply Gaussian noise, blurring and pixel intensity inversion. We detail the specific augmentations in B.3.

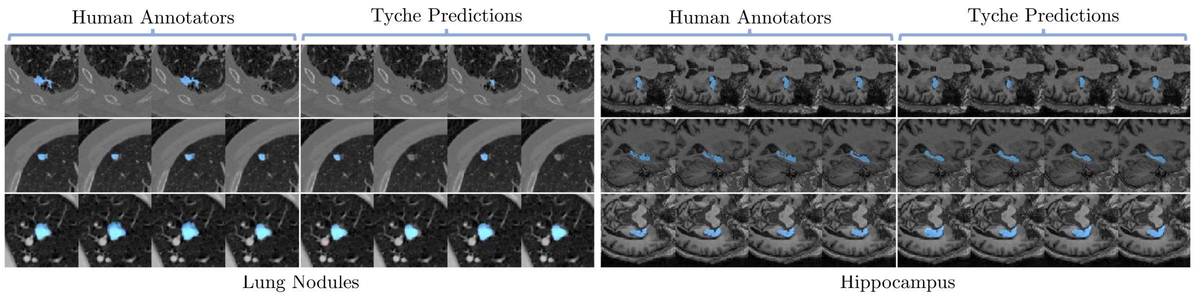

Figure 5. **Visualization of predictions for three different samples**, 1 per row. Left: LIDC-IDRI. Right: Hippocampus dataset. *Tyche* provides a set of prediction that is diverse and matches the raters, for tasks unseen at training time.

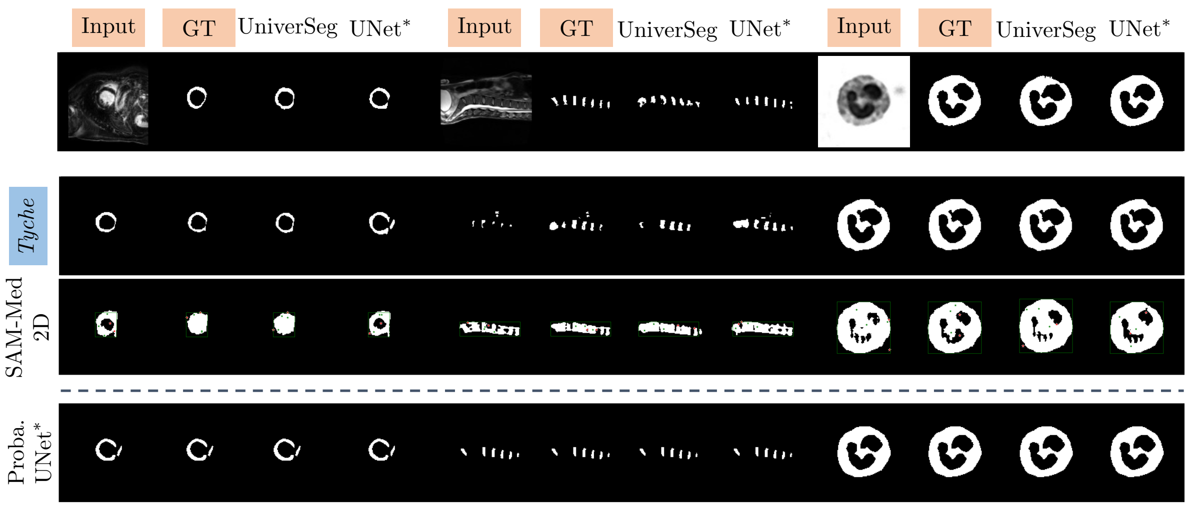

Figure 6. **Single annotator visualization for different models.** We show three example images that show very different corresponding segmentation. *Tyche* can output plausible segmentation for single annotator data with varying degrees of variability in the segmentation. Methods with an asterisk are upper baselines.

4. Experimental Setup

4.1. Data

We use a large collection of biomedical and synthetic datasets. Most datasets include a single manual segmentation per example, while a few have several raters per image.

Data Splits. We partition each dataset into development, validation, and test splits. We assign each dataset to an *in-distribution* set (*I.D.*) or an *out-of-distribution* set (*O.D.*). We train exclusively on the development splits of the *I.D.* datasets, and use the validation splits of the *I.D.* datasets to tune parameters. We use the validation splits of the *O.D.* datasets for final model selection. We report results on the test splits of the *O.D.* datasets. We find minimal difference between early stopping and training until convergence.

For each use case, we sample the context from each dataset's corresponding development set. Hence, the network doesn't see any of the *O.D.* datasets at training time.

Single-Annotator Data. For single annotator data, we build on MegaMedical used in recent publications [20, 134] and employ a collection of 73 public datasets, covering different biomedical domains [1, 3, 8, 15, 17, 18, 20, 22, 25, 36–38, 41, 42, 44, 45, 52, 56, 58, 59, 61, 71, 73, 74, 76, 77, 79–81, 84–91, 93, 98, 105, 106, 108, 112, 114, 117, 120, 123, 124, 138, 140–142]. MegaMedical spans a variety of anatomies and modalities, including brain MRI, cardiac ultrasound, thoracic CT and dental X-ray. We also use synthetic data involving simulated shapes, intensities, and image artifacts [20, 46]. The single-annotator datasets used for out-of-domain (*O.D.*) testing are: PanDental [1], WBC [142], SCD [106], ACDC [15], and SpineWeb [141].

Multi-Annotator Data. For multi-annotator *I.D.* data, we use four datasets from QUBIQ [94]: Brain Growth, Brain Lesions, Pancreas Lesions, and Kidney. We also simulate a multi-rater dataset consisting of random shapes (blobs). For the *O.D.* multi-annotator data we use four datasets. One contains hippocampus segmentation maps on brain MRIs from a large hospital. We crop the volumes around the hippocampus [69] to focus on the areas where the raters disagree. The second is a publicly available lung nodule dataset, LIDC-IDRI [5]. This dataset is notable for the substantial inter-rater variability. It contains 1,018 thoracic CT scans, each annotated by 4 annotators from a pool of 12 annotators. Finally, we also use retinal fundus images, STARE [49], annotated by 2 raters, and prostate data from the MICCAI 2021 QUBIQ challenge [94], annotated by 6 raters on two tasks. Single and multi-annotator combined, our *O.D.* group contains 20 tasks unseen at training time (some datasets have several tasks).

4.2. Evaluation

We evaluate our method by analysing individual prediction quality and distribution of predictions, both qualitatively and quantitatively. We also examine model choices through an ablation study.

A main use case of stochastic segmentation is to propose a small set of segmentations to a human rater, who can select the most appropriate one for their purpose. For this scenario, a model can be viewed as good if at least one prediction matches what the rater is looking for. We thus employ the best candidate Dice metric.

In the multi-annotator setting, we evaluate using both best candidate Dice score, also called maximum Dice score, as well as Generalized Energy Distance (GED) [14, 111, 125]. GED is commonly used in the stochastic segmentation literature to asses the difference between the distribution of predictions and the distribution of annotations [12, 69, 97, 107, 136]. GED has limitations, such as rewarding excessive prediction diversity [107]. Let $\mathcal{Y}$ and $\hat{\mathcal{Y}}$ be the set of annotations, GED is defined as:

$$D^2_{GED}(\mathcal{Y}, \hat{\mathcal{Y}}) = 2\mathbb{E}\left[d(p, \hat{p})\right] - \mathbb{E}\left[d(p, p')\right] - \mathbb{E}\left[d(\hat{p}, \hat{p}')\right], \quad (16)$$

where $p, p' \sim \mathcal{Y}$, $\hat{p}, \hat{p}' \sim \hat{\mathcal{Y}}$ and $d(\cdot, \cdot)$ is a distance metric. We use Dice score [31] as the distance metric.

4.3. Benchmarks

Tyche is the first method to produce stochastic segmentation predictions in-context. Consequently, we compare *Tyche* to existing benchmarks, each of which achieves only a subset of our goals.

In-Context Methods. We compare to deterministic frameworks that can leverage a context set: a few-shot method, SENet [110], and two in-context learning (ICL) methods, UniverSeg [20] and SegGPT [131]. We train UniverSeg and SENet with the same data splits and the same sets of augmentation transforms as for *Tyche*. For SegGPT, we use the public model, trained on a mix of natural and medical images. Figure 11 in the Supplemental Material shows that UniverSeg trained with additional data outperforms its public version.

	In-Context	Stochastic	Automatic
SENet	✓		✓
UniverSeg	✓		✓
SegGPT	✓		✓
Prob. UNet		✓	✓
PhiSeg		✓	✓
CIMD		✓	✓
SAM-based	✓	✓	
Tyche	✓	✓	✓

Table 1. **Summary of evaluated methods and their properties.** Only *Tyche* is both stochastic and in-context, and does not require user interaction.

Stochastic Upper Bounds. We compare to task-specialized probabilistic methods that are trained-on and perform well on specific datasets. We independently train Probabilistic UNet [69], PhiSeg[12] and CIDM, a recent diffusion network [107], on each of the 20 held-out tasks. For each task, we train three model variants: no augmentation, weak augmentation, and as much augmentation as for the *Tyche* targets. For each benchmark variant, we train on a *O.D.* development split and select the model that performs best on the corresponding *O.D.* validation split. We then compare these benchmarks to *Tyche* on the held-out *O.D.* test splits.

These models are explicitly optimized for the datasets on which they are evaluated, unlike *Tyche*, which does not use those datasets for training. Since these models are trained, tuned and evaluated on the *O.D.* datasets splits, something we explicitly aim to avoid in the problem set up as it is not easily done in many medical settings, they serve as upper bounds on performance.

Interactive Segmentation Methods. We compare to two interactive methods: SAM [68] and SAM-Med2D [24]. These methods can provide multiple segmentations, but, unlike *Tyche*, require human interaction, which is outside the scope of our work. SAM has a functionality to segment all elements in an image, however it is not optimized for medical imaging. We assume that the SAM-based models have access to the same information as the ICL methods: several image-segmentation pairs as context to guide the segmentation task. We fine-tune SAM using our *I.D.* development datasets. To replace the human interaction, we provide a bounding box, the average context label map, and 10 clicks, 5 positive and 5 negative. With SAM-Med2D, we use a bounding box, and 5 positive and negative clicks as input. For both SAM and SAM-Med2D, we generate clicks and bounding box from the average context label map.

We use one iteration of interaction, and sample different plausible segmentation candidates by sampling different sets of clicks and different averaged context sets.

Table 1 summarizes the features of all the methods. Ad-

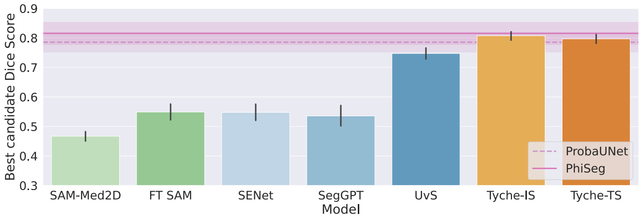

Figure 7. **Best candidate Dice score for single annotator data aggregated per task.** *Tyche* outperforms the in-context and interactive segmentation benchmarks, and approaches the stochastic upper bounds. Error bars represent 95% confidence intervals.

ditional information on the benchmarks is provided in the Supplemental Material.

4.4. Experiments

We evaluate all models on the multi-annotator and single-annotator *O.D.* data. We then analyze the *Tyche* variants individually and perform an ablation study on each to validate parameter choices. Finally, we compare the GPU inference runtimes and model parameters.

In the Supplemental Material, we analyze further the noise given as input, the context set, the number of predictions, the *SetBlock* and the candidate loss. We also provide additional performance metrics and per dataset results. We also compare the performance of *Tyche* and PhiSeg in a few-shot setting. Finally, we provide additional visualizations.

Inference Setting. We use a fixed context size of 16, because existing ICL systems show minimal improvements beyond this size [20]. Because there is variability in performance depending on the context sampled, we sample 5 different contexts for each datapoint and average performance. Similarly, for the stochastic upper bounds and interactive methods, we do 5 rounds of sampling K_i samples.

5. Results

5.1. Comparison to Benchmarks

Multi-Annotator O.D. Data. We evaluate on the datasets where *multiple annotations* exist for each sample. Figure 5 shows that, for both the lung nodules and the hippocampus datasets, *Tyche* predictions are diverse and capture rater diversity, even though these datasets are out-of-domain. Tables 2 and 3 show that both versions of *Tyche* outperform the interactive and deterministic benchmarks on all datasets except for Prostate Task 1, on which SegGPT has similar performance. Using a paired Student t-test, we find that *Tyche-TS* outperforms *Tyche-IS* in terms of best candidate Dice score, with $p < 10^{-10}$. We find no statistical difference between the two methods in terms of GED.

Single-Annotator Data. Figure 6 shows examples of predictions for *Tyche* and the corresponding benchmarks for the single-annotator datasets. *Tyche* produces a more diverse set of candidates than its competitors. Figure 7 compares all plausible models in terms of aggregate best candidate Dice score, except for CIDM, which underperformed for single-annotator data with a mean best candidate Dice score of: 0.673 ± 0.032. For clarity, we only present the full Figure in the Supplemental Material. *Tyche* performs better than the deterministic and interactive frameworks, and similarly to Probabilistic UNet, one of the upper bound benchmarks that is trained on the O.D. data. A paired Student t-test shows that *Tyche-IS* produces statistically higher GED ($p = 0.044$) than *Tyche-TS*, but we find no statistical difference in terms of best candidate Dice. We hypothesize that *Tyche-IS* is competitive because of the implicit annotator characterization provided by the context.

5.2. Tyche Analysis

We analyze *Tyche* variants and study the influence of different parameter choices.

Influence of the number of prediction K_i. We study how the number of predictions impacts the best candidate Dice score, keeping the context size constant. Figure 19 shows that for *Tyche-TS*, the best candidate Dice score rises with the number of predictions, but with diminishing returns.

Influence of context size $\|S\|$. Figure 21 shows that *Tyche-TS* is capable of leveraging the increased context size to improve its best candidate Dice score, and that a context size of 16 is sufficient to achieve most of the gain.

Ablation. Table 8 illustrates several ablations on *Tyche* design choices. For *Tyche-TS*, we evaluate the following variants: no simulated multi-annotator images, no *SetBlock* and finally, using the standard deviation of candidate feature representations in addition to the mean in the *SetBlock* ("Std"). We compare the models using best candidate Dice averaged across tasks.

Table 8 shows that the simulated multi-annotator data provides negligible improvement, as does adding the standard deviation. However, *SetBlock* is a crucial part to improve the best candidate Dice score.

We study performances for three types of TTA in *Tyche-IS*: on the target, on the context (CS), and on the context including the non-augmented context (CS+): $(S, \mathcal{G}(S))$. Table 8 shows that adding noise to only one of the target and context yields sub-optimal performance, while augmenting both the target and context improves performances.

5.3. Inference Runtime

We compare the inference runtime by predicting 8 segmentation candidates with each method, and repeat the process 300 times. We use an NVIDIA V100 GPU. Table 4 shows that *Tyche* is significantly faster and smaller than SegGPT

GED^2 (↓)		Hippocampus	LIDC-IDRI	Prostate Task 1	Prostate Task 2	STARE
Interactive	SAM	0.57 ± 0.02	0.90 ± 0.01	0.20 ± 0.03	0.31 ± 0.06	0.89 ± 0.06
	SAM-Med2d	0.93 ± 0.02	1.01 ± 0.01	0.80 ± 0.09	0.78 ± 0.11	1.52 ± 0.05
I-C & Stochastic (Ours)	**Tyche-IS**	$\mathbf{0.21 \pm 0.01}$	0.41 ± 0.01	0.12 ± 0.02	0.20 ± 0.05	0.73 ± 0.03
	Tyche-TS	0.22 ± 0.01	$\mathbf{0.40 \pm 0.01}$	$\mathbf{0.09 \pm 0.02}$	$\mathbf{0.15 \pm 0.03}$	$\mathbf{0.62 \pm 0.03}$
Stochastic Upper Bound	PhiSeg	0.14 ± 0.01	0.33 ± 0.01	0.12 ± 0.01	0.17 ± 0.05	1.22 ± 0.02
	ProbaUNet	0.13 ± 0.01	0.51 ± 0.01	0.08 ± 0.01	0.18 ± 0.05	0.76 ± 0.06
	CIDM	0.17 ± 0.01	0.42 ± 0.01	0.14 ± 0.02	0.26 ± 0.04	0.87 ± 0.05

Table 2. **Generalized Energy Distance** for different models with a context size of 16 for in-context methods and a number of predictions set to 8. Lower is better. *Tyche* outperforms interactive and ICL baselines, and matches stochastic upper bounds.

Max Dice (↑)		Hippocampus	LIDC-IDRI	Prostate Task 1	Prostate Task 2	STARE
In-Context	UniverSeg	0.84 ± 0.01	0.67 ± 0.01	0.91 ± 0.01	0.88 ± 0.03	0.51 ± 0.02
	SegGPT	0.10 ± 0.01	0.68 ± 0.01	0.94 ± 0.01	0.89 ± 0.03	0.02 ± 0.01
	SENet	0.68 ± 0.01	0.00 ± 0.00	0.83 ± 0.02	0.83 ± 0.02	0.30 ± 0.03
Interactive	SAM	0.71 ± 0.01	0.55 ± 0.01	0.90 ± 0.01	0.85 ± 0.03	0.50 ± 0.03
	SAM-Med2d	0.52 ± 0.01	0.42 ± 0.01	0.62 ± 0.04	0.64 ± 0.06	0.21 ± 0.03
I-C & Stochastic (Ours)	**Tyche-IS**	0.87 ± 0.01	0.90 ± 0.00	0.94 ± 0.01	0.91 ± 0.01	0.52 ± 0.03
	Tyche-TS	$\mathbf{0.88 \pm 0.01}$	$\mathbf{0.91 \pm 0.00}$	$\mathbf{0.95 \pm 0.01}$	$\mathbf{0.93 \pm 0.01}$	$\mathbf{0.60 \pm 0.02}$
Stochastic Upper Bound	PhiSeg	0.88 ± 0.00	0.91 ± 0.00	0.93 ± 0.01	0.91 ± 0.02	0.15 ± 0.01
	ProbaUNet	0.91 ± 0.00	0.86 ± 0.01	0.95 ± 0.00	0.91 ± 0.03	0.59 ± 0.02
	CIMD	0.84 ± 0.01	0.92 ± 0.00	0.93 ± 0.01	0.87 ± 0.02	0.41 ± 0.04

Table 3. **Best candidate Dice score** for different models with a context size of 16 for ICL methods and a number of predictions set to 8. Higher is better. *Tyche* outperforms interactive and ICL baselines, and matches stochastic upper bounds.

	Inference Time (ms)	Parameters
UniverSeg	96.62 ± 0.61	1.2M
SegGPT	$2{,}857.19 \pm 4.38$	370M
SENet	14.91 ± 0.21	0.89M
FT-SAM	$1{,}036.75 \pm 4.61$	94M
SAM-Med2D	188.8 ± 7.58	91M
PhiSeg	11.35 ± 0.672	21.1M
ProbaUNet	8.44 ± 0.46	5M
CIDM	$1.7 \times 10^5 \pm 2748$	85.6M
Tyche-IS	128.57 ± 2.626	1.2M
Tyche-TS	18.09 ± 0.61	1.7M

Table 4. **Inference Runtime and Model Parameters** for 8 predictions and a context size of 16.

and CIDM, yet, not as fast as some task-specific stochastic models. *Tyche-IS* has fewer parameters than *Tyche-TS*, but needs additional inference time.

6. Conclusion

We introduced *Tyche*, the first framework for stochastic in-context segmentation. For any (new) segmentation task, *Tyche* can directly produce diverse segmentation candidates, from which practitioners can select the most suitable one, and draw a better understanding of the underlying uncertainty. *Tyche* can generalize to images from datasets unseen at training and outperforms in-context and interactive benchmarks. In addition, *Tyche* often matches stochastic models on tasks for which those models have been specifically trained. *Tyche* has two variants, one designed to optimize the best segmentation candidate, with fast inference time, and a test-time augmentation variant that can be used in combination with existing in-context learning methods. We are excited to further study the different types of uncertainty captured by *Tyche-TS* and *Tyche-IS*.

7. Acknowledgement

This research was supported in part by the National Institute of Biomedical Imaging and Bioengineering of the National Institutes of Health under award number R01EB033773, the Eric and Wendy Schmidt Center at the Broad Institute of MIT and Harvard, Quanta Computer Inc as well as by CZI grant DAF2020-225720 (grant DOI https://doi.org/10.37921/977328pjvbca) from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation (funder DOI 10.13039/100014989). Some of the computation resources required for this research was performed on computational hardware generously provided by the Massachusetts Life Sciences Center.

References

[1] Amir Hossein Abdi, Shohreh Kasaei, and Mojdeh Mehdizadeh. Automatic segmentation of mandible in panoramic x-ray. *Journal of Medical Imaging*, 2(4):044003, 2015. 5, 18

[2] C. J. Aine, H. J. Bockholt, J. R. Bustillo, J. M. Cañive, A. Caprihan, C. Gasparovic, F. M. Hanlon, J. M. Houck, R. E. Jung, J. Lauriello, J. Liu, A. R. Mayer, N. I. Perrone-Bizzozero, S. Posse, J. M. Stephen, J. A. Turner, V. P. Clark, and Vince D. Calhoun. Multimodal Neuroimaging in Schizophrenia: Description and Dissemination. *Neuroinformatics*, 15(4):343–364, 2017. 1

[3] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020. 5, 18

[4] Mina Amiri, Rupert Brooks, Bahareh Behboodi, and Hassan Rivaz. Two-stage ultrasound image segmentation using u-net and test time augmentation. *International journal of computer assisted radiology and surgery*, 15:981–988, 2020. 3

[5] Samuel G Armato III and et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Medical physics*, 38(2):915–931, 2011. 6, 19

[6] 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. 3

[7] Vijay Badrinarayanan and et al. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 2

[8] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021. 5, 18

[9] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4 (1):1–13, 2017. 18

[10] Ivana Balažević and et al. Towards in-context scene understanding. *arXiv preprint arXiv:2306.01667*, 2023. 2

[11] Sophia Bano, Francisco Vasconcelos, Luke M Shepherd, Emmanuel Vander Poorten, Tom Vercauteren, Sebastien Ourselin, Anna L David, Jan Deprest, and Danail Stoyanov. Deep placental vessel segmentation for fetoscopic mosaicking. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, pages 763–773. Springer, 2020. 18

[12] Christian F Baumgartner, Kerem C Tezcan, Krishna Chaitanya, Andreas M Hötker, Urs J Muehlematter, Khoschy Schawkat, Anton S Becker, Olivio Donati, and Ender Konukoglu. Phiseg: Capturing uncertainty in medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 119–127. Springer, 2019. 2, 6

[13] Anton S Becker, Krishna Chaitanya, Khoschy Schawkat, Urs J Muehlematter, Andreas M Hötker, Ender Konukoglu, and Olivio F Donati. Variability of manual segmentation of the prostate in axial t2-weighted mri: a multi-reader study. *European journal of radiology*, 121:108716, 2019. 2

[14] Marc G Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The cramer distance as a solution to biased wasserstein gradients. *arXiv preprint arXiv:1705.10743*, 2017. 6

[15] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018. 5, 18

[16] Ishaan Bhat, Josien PW Pluim, and Hugo J Kuijf. Generalized probabilistic u-net for medical image segementation. In *International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*, pages 113–124. Springer, 2022. 2

[17] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019. 5, 18

[18] Nicholas Bloch, Anant Madabhushi, Henkjan Huisman, John Freymann, Justin Kirby, Michael Grauer, Andinet Enquobahrie, Carl Jaffe, Larry Clarke, and Keyvan Farahani. Nci-isbi 2013 challenge: automated segmentation of prostate structures. *The Cancer Imaging Archive*, 370(6):5, 2015. 5, 18

[19] Mateusz Buda, Ashirbani Saha, and Maciej A Mazurowski. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in biology and medicine*, 109: 218–225, 2019. 18

[20] Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Universeg: Universal medical image segmentation. *arXiv preprint arXiv:2304.06131*, 2023. 2, 3, 4, 5, 6, 7, 1

[21] Juan C. Caicedo, Allen Goodman, Kyle W. Karhohs, Beth A. Cimini, Jeanelle Ackerman, Marzieh Haghighi, CherKeng Heng, Tim Becker, Minh Doan, Claire McQuin, Mohammad Rohban, Shantanu Singh, and Anne E. Carpenter. Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. *Nature Methods*, 16(12):1247–1253, 2019. 18

[22] Albert Cardon, Stephan Saalfeld, Stephan Preibisch, Benjamin Schmid, Anchi Cheng, Jim Pulokas, Pavel Toman-

cak, and Volker Hartenstein. Isbi challenge: Segmentation of neuronal structures in em stacks. 5

[23] Guillaume Charpiat, Matthias Hofmann, and Bernhard Schölkopf. Automatic image colorization via multimodal predictions. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part III 10*, pages 126–139. Springer, 2008. 4

[24] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, Hui Sun, Junjun He, Shaoting Zhang, Min Zhu, and Yu Qiao. SAM-Med2D, 2023. arXiv:2308.16184 [cs]. 6

[25] Noel C. F. Codella, David A. Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin K. Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (ISIC). *CoRR*, abs/1710.05006, 2017. 5

[26] Seffi Cohen, Niv Goldshlager, Lior Rokach, and Bracha Shapira. Boosting anomaly detection using unsupervised diverse test-time augmentation. *Information Sciences*, 626: 821–836, 2023. 3

[27] Steffen Czolbe, Kasra Arnavaz, Oswin Krause, and Aasa Feragen. Is segmentation uncertainty useful? In *Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27*, pages 715–726. Springer, 2021. 2

[28] Etienne Decenciere, Guy Cazuguel, Xiwei Zhang, Guillaume Thibault, J-C Klein, Fernand Meyer, Beatriz Marcotegui, Gwénolé Quellec, Mathieu Lamard, Ronan Danno, et al. Teleophta: Machine learning and image processing methods for teleophthalmology. *Irbm*, 34(2):196–203, 2013. 18

[29] Aysen Degerli, Morteza Zabihi, Serkan Kiranyaz, Tahir Hamid, Rashid Mazhar, Ridha Hamila, and Moncef Gabbouj. Early detection of myocardial infarction in low-quality echocardiography. *IEEE Access*, 9:34442–34453, 2021. 18

[30] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009. 2

[31] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. 6

[32] Hao Ding, Changchang Sun, Hao Tang, Dawen Cai, and Yan Yan. Few-shot medical image segmentation with cycle-resemblance attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2488–2497, 2023. 2

[33] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 2

[34] Muhammad Moazam Fraz, Paolo Remagnino, Andreas Hoppe, Bunyarit Uyyanonvara, Alicja R. Rudnicka, Christopher G. Owen, and Sarah A. Barman. An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Transactions on Biomedical Engineering*, 59(9):2538–2548, 2012. 1

[35] Mélanie Gaillochet, Christian Desrosiers, and Hervé Lombaert. Taal: Test-time augmentation for active learning in medical image segmentation. In *Data Augmentation, Labelling, and Imperfections*. Springer Nature Switzerland, 2022. 3

[36] J Gamper, NA Koohbanani, K Benes, S Graham, M Jahanifar, SA Khurram, A Azam, K Hewitt, and N Rajpoot. Pannuke dataset extension, insights and baselines. arxiv. 2020 doi: 10.48550. *ARXIV*, 2003. 5

[37] Stephan Gerhard, Jan Funke, Julien Martel, Albert Cardona, and Richard Fetter. Segmented anisotropic ssTEM dataset of neural tissue. *figshare*, pages 0–0, 2013.

[38] Randy L Gollub, Jody M Shoemaker, Margaret D King, Tonya White, Stefan Ehrlich, Scott R Sponheim, Vincent P Clark, Jessica A Turner, Bryon A Mueller, Vince Magnotta, et al. The mcic collection: a shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia. *Neuroinformatics*, 11:367–388, 2013. 5, 18

[39] Ioannis S Gousias, Daniel Rueckert, Rolf A Heckemann, Leigh E Dyet, James P Boardman, A David Edwards, and Alexander Hammers. Automatic segmentation of brain mris of 2-year-olds into 83 regions of interest. *Neuroimage*, 40(2):672–684, 2008. 18

[40] Ioannis S Gousias, A David Edwards, Mary A Rutherford, Serena J Counsell, Jo V Hajnal, Daniel Rueckert, and Alexander Hammers. Magnetic resonance imaging of the newborn brain: manual segmentation of labelled atlases in term-born and preterm infants. *Neuroimage*, 62(3):1499–1509, 2012. 18

[41] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019. 5

[42] Daniel Gut. X-ray images of the hip joints. 1, 2021. Publisher: Mendeley Data. 5, 18

[43] Abner Guzman-Rivera, Dhruv Batra, and Pushmeet Kohli. Multiple choice learning: Learning to produce multiple structured outputs. *Advances in neural information processing systems*, 25, 2012. 4

[44] Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, page 101821, 2020. 5, 18

[45] Moritz R Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, et al. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific data*, 9(1):762, 2022. 5, 18

[46] Malte Hoffmann, Benjamin Billot, Douglas N Greve, Juan Eugenio Iglesias, Bruce Fischl, and Adrian V Dalca. Synthmorph: learning contrast-invariant registration without acquired images. *IEEE transactions on medical imaging*, 41(3):543–558, 2021. 5

[47] Sungmin Hong, Anna K Bonkhoff, Andrew Hoopes, Martin Bretzner, Markus D Schirmer, Anne-Katrin Giese, Adrian V Dalca, Polina Golland, and Natalia S Rost. Hypernet-ensemble learning of segmentation probability for medical image segmentation with ambiguous labels. *arXiv preprint arXiv:2112.06693*, 2021. 2

[48] Andrew Hoopes, Malte Hoffmann, Douglas N. Greve, Bruce Fischl, John Guttag, and Adrian V. Dalca. Learning the effect of registration hyperparameters with hypermorph. pages 1–30, 2022. 18

[49] AD Hoover, Valentina Kouznetsova, and Michael Goldbaum. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical imaging*, 19(3):203–210, 2000. 6, 19

[50] Qingqiao Hu, Hao Wang, Jing Luo, Yunhao Luo, Zhiheng Zhangg, Jan S Kirschke, Benedikt Wiestler, Bjoern Menze, Jianguo Zhang, and Hongwei Bran Li. Inter-rater uncertainty quantification in medical image segmentation via rater-specific bayesian neural networks. *arXiv preprint arXiv:2306.16556*, 2023. 2

[51] Xiaoqiong Huang, Zejian Chen, Xin Yang, Zhendong Liu, Yuxin Zou, Mingyuan Luo, Wufeng Xue, and Dong Ni. Style-invariant cardiac image segmentation with test-time augmentation. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers 11*, pages 305–315. Springer, 2021. 3

[52] Humans in the Loop. Teeth segmentation dataset. 5, 18

[53] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 2

[54] Mobarakol Islam and Ben Glocker. Spatially varying label smoothing: Capturing uncertainty from expert annotations. In *Information Processing in Medical Imaging: 27th International Conference, IPMI 2021, Virtual Event, June 28–June 30, 2021, Proceedings 27*, pages 677–688. Springer, 2021. 2

[55] Debesh Jha, Pia H Smedsrud, Dag Johansen, Thomas de Lange, Håvard D Johansen, Pål Halvorsen, and Michael A Riegler. A comprehensive study on colorectal polyp segmentation with resunet++, conditional random field and test-time augmentation. *IEEE journal of biomedical and health informatics*, 25(6):2029–2040, 2021. 3

[56] Yuanfeng Ji, Haotian Bai, Jie Yang, Chongjian Ge, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *arXiv preprint arXiv:2206.08023*, 2022. 5, 18

[57] Leo Joskowicz, D Cohen, N Caplan, and Jacob Sosna. Inter-observer variability of manual contour delineation of structures in ct. *European radiology*, 29:1391–1399, 2019. 2

[58] Rashed Karim, R James Housden, Mayuragoban Balasubramaniam, Zhong Chen, Daniel Perry, Ayesha Uddin, Yosra Al-Beyatti, Ebrahim Palkhi, Prince Acheampong, Samantha Obom, et al. Evaluation of current algorithms for segmentation of scar tissue from late gadolinium enhancement cardiovascular magnetic resonance of the left atrium: an open-access grand challenge. *Journal of Cardiovascular Magnetic Resonance*, 15(1):1–17, 2013. 5, 18

[59] Ali Emre Kavur, M. Alper Selver, Oğuz Dicle, Mustafa Barış, and N. Sinem Gezer. CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data. 2019. 5

[60] Ali Emre Kavur, M. Alper Selver, Oğuz Dicle, Mustafa Barış, and N. Sinem Gezer. CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data, 2019. 18

[61] A. Emre Kavur, N. Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, Bora Baydar, Dmitry Lachinov, Shuo Han, Josef Pauli, Fabian Isensee, Matthias Perkonigg, Rachana Sathish, Ronnie Rajan, Debdoot Sheet, Gurbandurdy Dovletov, Oliver Speck, Andreas Nürnberger, Klaus H. Maier-Hein, Gözde Bozdağı Akar, Gözde Ünal, Oğuz Dicle, and M. Alper Selver. CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021. 5

[62] A. Emre Kavur, N. Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, Bora Baydar, Dmitry Lachinov, Shuo Han, Josef Pauli, Fabian Isensee, Matthias Perkonigg, Rachana Sathish, Ronnie Rajan, Debdoot Sheet, Gurbandurdy Dovletov, Oliver Speck, Andreas Nürnberger, Klaus H. Maier-Hein, Gözde Bozdağı Akar, Gözde Ünal, Oğuz Dicle, and M. Alper Selver. CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021. 18

[63] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 2

[64] 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. 2

[65] Donggyun Kim, Jinwoo Kim, Seongwoong Cho, Chong Luo, and Seunghoon Hong. Universal few-shot learning of dense prediction tasks with visual token matching. *arXiv preprint arXiv:2303.14969*, 2023. 2

[66] Kwanyoung Kim, Dongwon Park, Kwang In Kim, and Se Young Chun. Task-aware variational adversarial active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8166–8175, 2021. 3

[67] Serkan Kiranyaz, Aysen Degerli, Tahir Hamid, Rashid Mazhar, Rayyan El Fadil Ahmed, Rayaan Abouhasera, Morteza Zabihi, Junaid Malik, Ridha Hamila, and Moncef Gabbouj. Left ventricular wall motion estimation by active polynomials for acute myocardial infarction detection. *IEEE Access*, 8:210301–210317, 2020. 18

[68] Alexander Kirillov and et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 4, 6

[69] Simon Kohl, Bernardino Romera-Paredes, Clemens Meyer, Jeffrey De Fauw, Joseph R Ledsam, Klaus Maier-Hein, SM Eslami, Danilo Jimenez Rezende, and Olaf Ronneberger. A probabilistic u-net for segmentation of ambiguous images. *Advances in neural information processing systems*, 31, 2018. 2, 6

[70] Simon AA Kohl, Bernardino Romera-Paredes, Klaus H Maier-Hein, Danilo Jimenez Rezende, SM Eslami, Pushmeet Kohli, Andrew Zisserman, and Olaf Ronneberger. A hierarchical probabilistic u-net for modeling multi-scale ambiguities. *arXiv preprint arXiv:1905.13077*, 2019. 2, 9

[71] Markus Krönke, Christine Eilers, Desislava Dimova, Melanie Köhler, Gabriel Buschner, Lilit Schweiger, Lemonia Konstantinidou, Marcus Makowski, James Nagarajah, Nassir Navab, et al. Tracked 3d ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry. *Plos one*, 17(7):e0268550, 2022. 5

[72] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 9

[73] Hugo J Kuijf, J Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adria Casamitjana, et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE transactions on medical imaging*, 38(11):2556–2568, 2019. 5, 18

[74] Maria Kuklisova-Murgasova, Paul Aljabar, Latha Srinivasan, Serena J Counsell, Valentina Doria, Ahmed Serag, Ioannis S Gousias, James P Boardman, Mary A Rutherford, A David Edwards, et al. A dynamic 4d probabilistic atlas of the developing brain. *NeuroImage*, 54(4):2750–2763, 2011. 5, 18

[75] Benjamin Lambert, Florence Forbes, Senan Doyle, and Michel Dojat. Triadnet: Sampling-free predictive intervals for lesional volume in 3d brain mr images. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging – MICCAI 2023*. Springer Nature Switzerland, 2023. 2

[76] Zoé Lambert, Caroline Petitjean, Bernard Dubray, and Su Kuan. Segthor: segmentation of thoracic organs at risk in ct images. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2020. 5, 18

[77] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, page 12, 2015. 5, 18

[78] Agostina Larrazabal, Cesar Martinez, Jose Dolz, and Enzo Ferrante. Maximum entropy on erroneous predictions (meep): Improving model calibration for medical image segmentation. *arXiv preprint arXiv:2112.12218*, 2021. 2

[79] Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019. 5, 18

[80] Guillaume Lemaître, Robert Martí, Jordi Freixenet, Joan C Vilanova, Paul M Walker, and Fabrice Meriaudeau. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review. *Computers in biology and medicine*, 60:8–31, 2015. 18

[81] Mingchao Li, Yuhan Zhang, Zexuan Ji, Keren Xie, Songtao Yuan, Qinghuai Liu, and Qiang Chen. Ipn-v2 and octa-500: Methodology and dataset for retinal image segmentation. *arXiv preprint arXiv:2012.07261*, 2020. 5, 18

[82] Yiwen Li, Yunguan Fu, Iani Gayo, Qianye Yang, Zhe Min, Shaheer Saeed, Wen Yan, Yipei Wang, J Alison Noble, Mark Emberton, et al. Prototypical few-shot segmentation for cross-institution male pelvic structures with spatial registration. *arXiv preprint arXiv:2209.05160*, 2022. 2

[83] Zhuwen Li, Qifeng Chen, and Vladlen Koltun. Interactive image segmentation with latent diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 577–585, 2018. 4

[84] Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014. 5, 18

[85] Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature methods*, 9(7):637–637, 2012. 18

[86] Maximilian T Löffler, Anjany Sekuboyina, Alina Jacob, Anna-Lena Grau, Andreas Scharr, Malek El Husseini, Mareike Kallweit, Claus Zimmer, Thomas Baum, and Jan S Kirschke. A vertebral segmentation dataset with fracture grading. *Radiology: Artificial Intelligence*, 2(4):e190138, 2020.

[87] Xiangde Luo, Wenjun Liao, Jianghong Xiao, Tao Song, Xiaofan Zhang, Kang Li, Guotai Wang, and Shaoting Zhang. Word: Revisiting organs segmentation in the whole abdominal region. *arXiv preprint arXiv:2111.02403*, 2021. 18

[88] Yuhui Ma, Huaying Hao, Jianyang Xie, Huazhu Fu, Jiong Zhang, Jianlong Yang, Zhen Wang, Jiang Liu, Yalin Zheng, and Yitian Zhao. Rose: a retinal oct-angiography vessel segmentation dataset and new model. *IEEE Transactions on Medical Imaging*, 40(3):928–939, 2021. 18

[89] Jacob A. Macdonald, Zhe Zhu, Brandon Konkel, Maciej Mazurowski, Walter Wiggins, and Mustafa Bashir. Duke liver dataset (MRI) v2, 2023.

[90] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open

access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9): 1498–1507, 2007. 18

[91] Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, et al. The parkinson progression marker initiative (ppmi). *Progress in neurobiology*, 95(4):629–635, 2011. 5, 18

[92] Kazuhisa Matsunaga, Akira Hamada, Akane Minagawa, and Hiroshi Koga. Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble. *arXiv preprint arXiv:1703.03108*, 2017. 3

[93] Maciej A Mazurowski, Kal Clark, Nicholas M Czarnek, Parisa Shamsesfandabadi, Katherine B Peters, and Ashirbani Saha. Radiogenomics of lower-grade glioma: algorithmically-assessed tumor shape is associated with tumor genomic subtypes and patient outcomes in a multi-institutional study with the cancer genome atlas data. *Journal of neuro-oncology*, 133:27–35, 2017. 5, 18

[94] Bjoern Menze, Leo Joskowicz, Spyridon Bakas, Andras Jakab, Ender Konukoglu, Anton Becker, Amber Simpson, and Richard D. Quantification of uncertainties in biomedical image quantification 2021. *4th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2021)*, 2021. 5, 6, 19

[95] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 18

[96] 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. 4

[97] Miguel Monteiro, Loïc Le Folgoc, Daniel Coelho de Castro, Nick Pawlowski, Bernardo Marques, Konstantinos Kamnitsas, Mark van der Wilk, and Ben Glocker. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *Advances in Neural Information Processing Systems*, 33:12756–12767, 2020. 2, 6, 9

[98] Anna Montoya, Hasnin, kaggle446, shirzad, Will Cukierski, and yffud. Ultrasound nerve segmentation, 2016. 5

[99] Nikita Moshkov, Botond Mathe, Attila Kertesz-Farkas, Reka Hollandi, and Peter Horvath. Test-time augmentation for deep learning-based cell segmentation on microscopy images. *Scientific reports*, 10(1):5068, 2020. 3

[100] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 622–631, 2019. 2

[101] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 2

[102] Brennan Nichyporuk, Jillian Cardinell, Justin Szeto, Raghav Mehta, Jean-Pierre R Falet, Douglas L Arnold, Sotirios A Tsaftaris, and Tal Arbel. Rethinking generalization: The impact of annotation style on medical image segmentation. *arXiv preprint arXiv:2210.17398*, 2022. 2

[103] Prashant Pandey, Mustafa Chasmai, Tanuj Sur, and Brejesh Lall. Robust prototypical few-shot organ segmentation with regularized neural-odes. *IEEE Transactions on Medical Imaging*, 2023. 2

[104] Kelly Payette, Priscille de Dumast, Hamza Kebiri, Ivan Ezhov, Johannes C Paetzold, Suprosanna Shit, Asim Iqbal, Romesa Khan, Raimund Kottke, Patrice Grehten, et al. An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. *Scientific Data*, 8(1):1–14, 2021. 18

[105] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid), 2018. 5, 18

[106] Perry Radau, Yingli Lu, Kim Connelly, Gideon Paul, AJWG Dick, and Graham Wright. Evaluation framework for algorithms segmenting short axis cardiac mri. *The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge*, 49, 2009. 5, 18

[107] Aimon Rahman, Jeya Maria Jose Valanarasu, Ilker Hacihaliloglu, and Vishal M Patel. Ambiguous medical image segmentation using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11536–11546, 2023. 2, 6

[108] Blaine Rister, Darvin Yi, Kaushik Shivakumar, Tomomi Nobashi, and Daniel L. Rubin. CT-ORG, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7(1):381, 2020. 5, 18

[109] 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*. Springer International Publishing, 2015. 2, 3

[110] Abhijit Guha Roy, Shayan Siddiqui, Sebastian Pölsterl, Nassir Navab, and Christian Wachinger. Squeeze & excite guided few-shot segmentation of volumetric images. *Medical image analysis*, 59:101587, 2020. 6

[111] Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving gans using optimal transport. *arXiv preprint arXiv:1803.05573*, 2018. 6

[112] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, SQ Truong, CD Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, AY Ng, et al. Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation. *MedRxiv*, 2021. 5, 18

[113] Arne Schmidt, Pablo Morales-Álvarez, and Rafael Molina. Probabilistic modeling of inter-and intra-observer variability in medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21097–21106, 2023. 2

[114] Constantin Seibold, Simon Reiß, Saquib Sarfraz, Matthias A. Fink, Victoria Mayer, Jan Sellner, Moon Sung Kim, Klaus H. Maier-Hein, Jens Kleesiek, and Rainer

Stiefelhagen. Detailed annotations of chest x-rays via ct projection for report understanding. In *Proceedings of the 33th British Machine Vision Conference (BMVC)*, 2022. 5, 18

[115] Jun Seo, Young-Hyun Park, Sung Whan Yoon, and Jaekyun Moon. Task-adaptive feature transformer with semantic enrichment for few-shot segmentation. *arXiv preprint arXiv:2202.06498*, 2022. 2

[116] Ahmed Serag, Paul Aljabar, Gareth Ball, Serena J Counsell, James P Boardman, Mary A Rutherford, A David Edwards, Joseph V Hajnal, and Daniel Rueckert. Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression. *Neuroimage*, 59(3):2255–2265, 2012. 18

[117] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42:1–13, 2017. 5, 18

[118] Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. Better aggregation in test-time augmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1214–1223, 2021. 3

[119] Qianqian Shen, Yanan Li, Jiyong Jin, and Bin Liu. Q-net: Query-informed few-shot medical image segmentation. *arXiv preprint arXiv:2208.11451*, 2022. 2

[120] 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. 5, 18

[121] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5972–5981, 2019. 3

[122] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 2

[123] Yuxin Song, Jing Zheng, Long Lei, Zhipeng Ni, Baoliang Zhao, and Ying Hu. CT2US: Cross-modal transfer learning for kidney segmentation in ultrasound images with synthesized data. *Ultrasonics*, 122:106706, 2022. 5

[124] Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4):501–509, 2004. 5, 18

[125] Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013. 6

[126] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11244–11253, 2019. 2

[127] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016. 2

[128] 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. 3

[129] Guotai Wang, Wenqi Li, Sébastien Ourselin, and Tom Vercauteren. Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*, pages 61–72. Springer, 2019. 3

[130] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023. 2

[131] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Towards segmenting everything in context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1130–1140, 2023. 2, 6

[132] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5), 2023. 1

[133] Julia Wolleb, Robin Sandkühler, Florentin Bieder, Philippe Valmaggia, and Philippe C. Cattin. Diffusion Models for Implicit Image Segmentation Ensembles. In *Medical Imaging with Deep Learning*, 2021. 2

[134] Hallee E Wong, Marianne Rakic, John Guttag, and Adrian V Dalca. Scribbleprompt: Fast and flexible interactive segmentation for any medical image. *arXiv preprint arXiv:2312.07381*, 2023. 5, 1

[135] Andre Ye, Quan Ze Chen, and Amy Zhang. Confidence contours: Uncertainty-aware annotation for medical semantic segmentation. *arXiv preprint arXiv:2308.07528*, 2023. 2

[136] Lukas Zbinden, Lars Doorenbos, Theodoros Pissas, Adrian Thomas Huber, Raphael Sznitman, and Pablo Márquez-Neila. Stochastic segmentation with conditional categorical diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1119–1129, 2023. 2, 6, 9

[137] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5217–5226, 2019. 2

[138] Yingtao Zhang, Min Xian, Heng-Da Cheng, Bryar Shareef, Jianrui Ding, Fei Xu, Kuan Huang, Boyu Zhang, Chunping Ning, and Ying Wang. Busis: A benchmark for breast ultrasound image segmentation. In *Healthcare*, page 729. MDPI, 2022. 5

[139] Yingtao Zhang, Min Xian, Heng-Da Cheng, Bryar Shareef, Jianrui Ding, Fei Xu, Kuan Huang, Boyu Zhang, Chunping Ning, and Ying Wang. Busis: A benchmark for breast ultrasound image segmentation. In *Healthcare*, page 729. MDPI, 2022. 18

[140] Qi Zhao, Shuchang Lyu, Wenpei Bai, Linghan Cai, Binghao Liu, Meijing Wu, Xiubo Sang, Min Yang, and Lijiang Chen. A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation. *CoRR*, abs/2207.06799, 2022. 5

[141] Guoyan Zheng, Chengwen Chu, Daniel L Belavỳ, Bulat Ibragimov, Robert Korez, Tomaž Vrtovec, Hugo Hutt, Richard Everson, Judith Meakin, Isabel Lŏpez Andrade, et al. Evaluation and comparison of 3d intervertebral disc localization and segmentation methods for 3d t2 mr data: A grand challenge. *Medical image analysis*, 35:327–344, 2017. 5, 18

[142] Xin Zheng, Yong Wang, Guoyou Wang, and Jianguo Liu. Fast and robust segmentation of white blood cell images by self-supervised learning. *Micron*, 107:55–71, 2018. 5, 18