Pancakes: Consistent Multi-Protocol Image Segmentation Across Biomedical Domains

Marianne Rakic
MIT CSAIL, MGH
mrakic@mit.edu

Siyu Gai
MIT CSAIL, MGH

Etienne Chollet
MIT CSAIL, MGH

John V. Guttag
MIT CSAIL

Adrian V. Dalca
MIT CSAIL, HMS, MGH

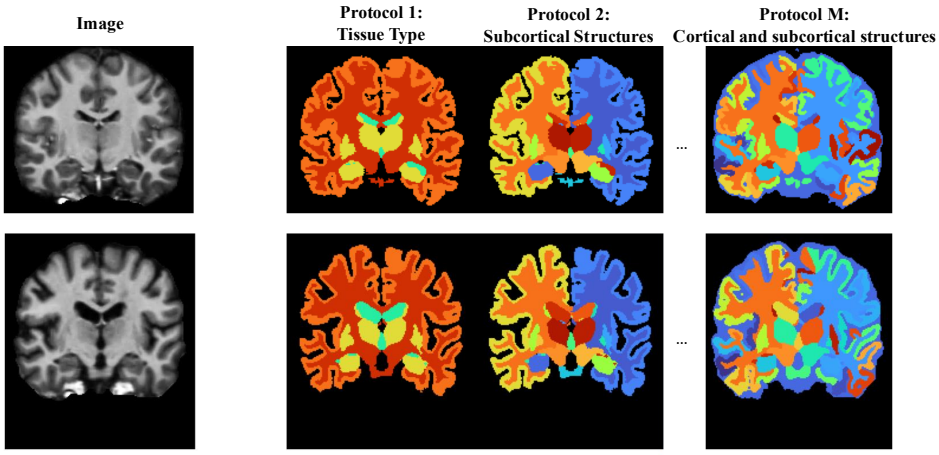


Figure 1: **Examples of different expert provided protocols for brain MRI.** Any biomedical image can be segmented in many different ways. For example, protocol 1 here corresponds to a coarse-grained categorization of tissue types. Colors correspond to distinct ROIs (the choice of colors is arbitrary). Typical neural networks follow a *fixed* protocol, specified explicitly or implicitly by the user.

Abstract

A single biomedical image can be meaningfully segmented in multiple ways, depending on the desired application. For instance, a brain MRI can be segmented according to tissue types, vascular territories, broad anatomical regions, fine-grained anatomy, or pathology, etc. Existing automatic segmentation models typically either (1) support only a single protocol – the one they were trained on – or (2) require labor-intensive manual prompting to specify the desired segmentation. We introduce Pancakes, a framework that, given a new image from a previously unseen domain, automatically generates multi-label segmentation maps for *multiple* plausible protocols, while maintaining semantic consistency across related images. Pancakes introduces a new problem formulation that is not currently attainable by existing foundation models. In a series of experiments on seven held-out datasets, we demonstrate that our model can significantly outperform existing foundation models in producing several plausible whole-image segmentations, that are semantically coherent across images.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

1 Introduction

There are many ways to segment a biomedical image. Depending on their goals, clinicians or biomedical researchers employ a specific segmentation *protocol*, which defines the regions of interest (ROIs) to be segmented (Figure 1). This could involve, for example, segmenting major anatomical classes, granular anatomical regions, diffuse tissue types, systemic structures like vessels or nerves, pathologies, or functional areas [6, 28, 45, 63, 80, 119].

Existing learning-based biomedical image segmentation tools require specification of a segmentation protocol [16, 58, 116, 123, 130]. Fully-automated models are trained to segment an image using image-segmentation training pairs, which implicitly define the protocol [18, 48, 99]. Recent in-context or few-shot models also use image-segmentation pairs, but take them *as input* to specify the desired segmentation protocol for a target image [8, 16, 96, 117]. Interactive models rely on user interactions to indicate the desired ROIs [58, 116, 122, 123]. In all of these strategies, the desired segmentation protocol is specified by the user (interactively or by example). This is a substantial burden on biomedical researchers, who commonly need to segment a new biomedical image with a potentially new segmentation protocol [4, 16].

Our goal is to support a *new capability*: enabling biomedical researchers and clinicians to explore a diverse set of plausible, semantically consistent segmentations for a previously unseen collection of images. We propose a fundamentally new approach to segmenting a new biomedical dataset. Instead of requiring the user to specify the protocol for a new task, our method, Pancakes, produces segmentation maps for multiple plausible protocols simultaneously, each consistent across images. After Pancakes has generated the label maps, a researcher or clinician can select which of the proposed protocols best aligns with their intended downstream use (e.g., anatomical volume analysis). We envision at least two broad classes of use:

(1) Rapid segmentation for new protocols. New protocols are frequently introduced [1, 36, 85], and there is a need to produce corresponding segmentations. If a scientist has a particular protocol in mind, but there are no existing tools for segmenting it, they can choose the protocol from Pancakes that best aligns with their intended use.

(2) Exploratory population analysis. Pancakes will support users in discovering or selecting segmentation strategies appropriate for their scientific or clinical questions. For example, a clinical scientist who studies how anatomy relates to some outcome (e.g., progression of a disease) or predictor (e.g., genetics) can use Pancakes to quickly extract segmentations of multiple candidate anatomical regions that have never been segmented, compute their volumes, and test correlations with clinical outcomes thereby identifying promising candidate regions.

Pancakes takes an image as input and produces a *distribution* over segmentation protocols. It then uses a segmentation sampling mechanism to produce several complete, multi-label segmentation maps from diverse protocols for that image. Importantly, within a chosen protocol, segmentation maps are semantically consistent across subjects – a specific label denotes the same anatomical structure in every image of the collection.

In a series of experiments, we demonstrate that our model can significantly outperform baselines in producing several plausible whole-image segmentations that are semantically coherent across images. We show that Pancakes outperforms foundation segmentation models by a wide margin on seven held-out datasets.

2 Related work

Single-protocol biomedical image segmentation. Most existing biomedical image segmentation models [18, 48, 99] are by design constrained to a specific biomedical domain and image type. For example, some models specialize in segmenting images of brains [13, 28, 38, 77, 92], hearts [112, 134], or eye vessels [50, 70, 100]. Each model learns a specific segmentation protocol defined by the image-segmentation pairs used during training.

Universal segmentation. Recent *universal* models can each segment a wide variety of structures across biomedical domains. They are trained jointly on large data collections containing diverse

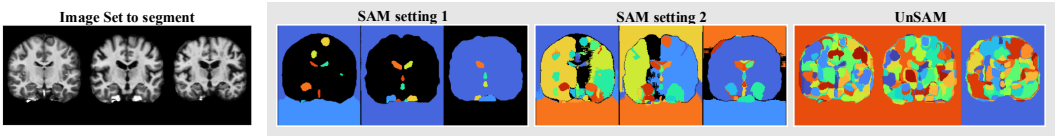


Figure 2: **Current automatic segmentation foundation models produce inconsistent segmentations.** Given a set of similar images to segment (left), automatic foundation models can fully parcellate each image, but the obtained segmentations are not semantically consistent across images. Even in rare cases where the same structure is labeled on two images, the *label index* (color-coded here) is usually inconsistent. There is then no clear mapping between the segmentations from the two samples, making it difficult for biomedical clinicians and researchers to use the results.

structures and image types, both for natural [8, 58, 97, 115, 117] and medical imaging [16, 44, 96, 123, 124, 130, 131]. Some methods generalize to new tasks by enabling a condition, or prompt, as input. This conditioning could involve example image-segmentation pairs [8, 16, 30, 96, 115, 117, 122], user interactions such as bounding boxes, clicks, or scribbles [58, 76, 97, 122–124, 135], or even text [44, 130, 131, 135]. Providing this conditioning is labor-intensive, especially when tackling a new segmentation task involving a large collection of images.

We also train Pancakes using large image collections, to generalize to new domains. However, we avoid the need for the user to laboriously prescribe the protocol and instead automatically estimate several segmentation maps from several plausible protocols.

Some universal models can completely partition an image from a new biomedical domain into multiple labels [58, 68, 116]. As we show in our experiments and illustrate in Figure 2, these segmentation maps are semantically inconsistent across subjects, with the same label having different meanings across images. In the rare case when the same structure is segmented in two images, the assigned label index is usually different, and establishing the correspondence is non-trivial.

Multi-protocol segmentation. Motivated by the fact that many objects in an image can be divided into subparts, some methods produce labels from a hierarchy of protocols in an image [24, 88, 114, 125]. This restricts the types of segmentations produced to fixed protocols that are inherently hierarchical. The methods are trained on limited domains or specialized to natural images. In either case, they require the hierarchy to be explicitly provided, which makes them less broadly applicable. In contrast, our approach produces label maps from multiple protocols that need not be hierarchically related, while generalizing to unseen structures.

Ambiguity and uncertainty. Even within a well-defined protocol, many segmentation tasks and biomedical images involve substantial ambiguity. This can be caused by problems with the image acquisition (e.g., noise or low contrast), ambiguous definitions of the desired region, or the downstream goals following the segmentation step. Recent models capture variability among manual raters [96, 102, 111], often by aggregating multiple predictions for a given structure or protocol to obtain an uncertainty estimate [23]. In our work, we jointly capture the ambiguity of the possible protocol and the inherent ambiguity in the image, but focus on the ability of a single framework to produce segmentations that represent different protocols consistently across scans.

Deep-learning and sampling mechanisms. Deep-learning segmentation frameworks that produce different outputs for a given input use an implicit or explicit mechanism to sample different solutions, such as variational autoencoders [10, 59, 60], diffusion models [95, 120, 121, 126], multivariate Gaussian [86], or in-context stochastic models [96]. We build on these methods, and propose a new mechanism to sample different segmentation protocols that are varied but consistent among images from the same domain.

3 Method

Given an image x, we let y_m be a multi-label segmentation map for a specific protocol m composed of non-overlapping labels. Typically, a segmentation model $g_\theta(x) = \hat{y}_m$ follows a fixed predefined protocol m.

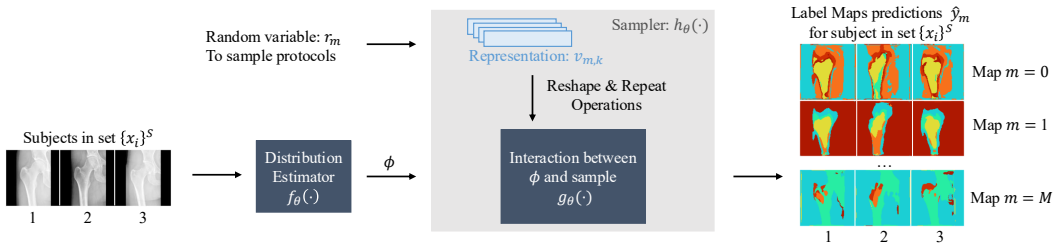


Figure 3: **Method Schematic.** To produce multiple consistent label maps for an image set $\{x_s\}$, we first estimate the distribution parameters ϕ via $f_{\theta_f}(x)$. We then sample from the distribution with parameters ϕ through the random variable r_m: $h_{\theta_h}(r_m, \phi) = \hat{y}_m$.

Instead, we design a framework that can produce a set of label maps $\{y_m\}_{m=1}^M$ for M different protocols, summarized in Figure 3. We estimate high-dimensional parameters ϕ of a distribution $p(y_m; \phi|x)$ over segmentation maps y_m spanning different segmentation protocols m, using the function $f_{\theta_f}(x) = \phi$. We model the distribution parameters ϕ as a vector for every image pixel, encoding the likelihood of different labels at that pixel location across protocols.

We model $f_{\theta_f}(x) = \phi$ as a neural network with a UNet architecture, which takes in an image and outputs the parameters ϕ. Below, we describe the mechanism for sampling segmentation maps from the distribution $p(y_m; \phi|x)$. We then define a loss that encourages segmentation maps for a given protocol m to be semantically consistent across a set of S images from the same biomedical domain.

Protocol sampling. We design a new mechanism to produce diverse segmentation maps across different protocols $y_m \sim p(y_m; \phi|x)$. Let a random integer $r_m = (M, K) \sim \mathcal{U}(1, M^{\text{max}}) \times \mathcal{U}(1, K^{\text{max}})$, where $\mathcal{U}$ is the discrete uniform distribution, M^{max} is a maximum number of label maps and K^{max} a maximum number of labels for per map. We compute label map $h_{\theta_h}(\phi, r_m) = \hat{y}_m$ given a deterministic function h_{θ_h}. The function first forms an intermediate representation $v_m = e(r_m)$, building on concepts from positional embedding. We concatenate this representation with the distribution parameters at each image location, and use a shallow fully-convolutional network to yield the final segmentation maps:

$$\hat{y}_m = h_{\theta_h}(\phi || v_m) \tag{1}$$

$$= h_{\theta_h}(f_{\theta_f}(x) || e(r_m)). \tag{2}$$

We predict the distributional parameters ϕ once, and then efficiently compute $h_{\theta_h}(\phi || e(\{r_m\})) = \{\hat{y}_m\}$ for a set of random integers $\{r_m\}$.

The intermediate representations $v_m = \{v_{m,k}\}$ capture all labels k in protocol m. Let u_m and u_k be vector representations corresponding to protocol m and label k, inspired by position embedding [113]. We model $v_{m,k} = u_m || u_k$, where $||$ denotes the concatenation operation. Specifically, given integer values t, we form vector representation u_t as:

$$u_t^{2j} = \sin(z_{t,2j} + \frac{\pi}{2}), \qquad u_t^{2j+1} = \sin(z_{t,2j} - \frac{\pi}{2}) \tag{3}$$

with

$$z_{t,j} = \frac{t\,\pi}{T} 2^{\frac{2j\,\pi}{J}}, \tag{4}$$

where u_t^j denotes entry j in vector u_t, $j = 1, \ldots, J$ and J is a hyperparameter determining the size of the vector. We use this formulation to form vectors for any desired u_m and u_k, which, in turn, enables us to form v_m for any given protocol. Specifically, the periods T for both vector representations u_m and u_k are determined by the random variable r_m, i.e. the values of M and K sampled.

Inference. During inference, Pancakes is given a set of images $\{x_s\}$ as input. For each image in the set, Pancakes produces a variable number of label maps $\hat{y}_{s,m,k}$ from various protocols with a variable number of labels k semantically consistent across the images.

3.1 Training strategy

Our goal is to enable off-the-shelf multi-protocol segmentation of *any* biomedical image x, especially for those not seen during training. To achieve this, we train a single Pancakes model on a wide

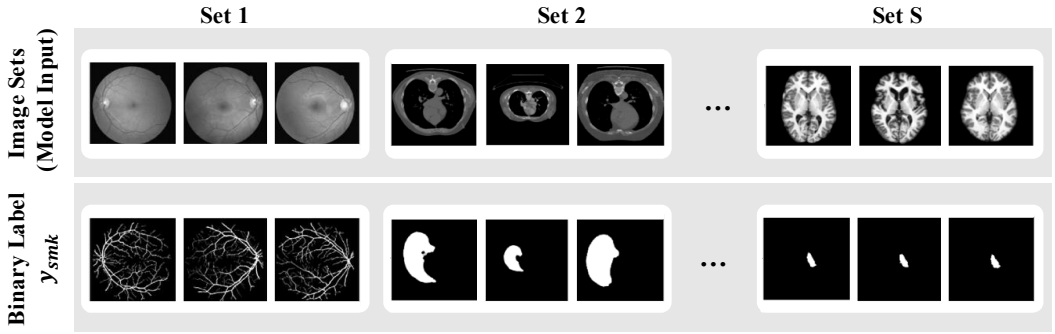


Figure 4: **Example input and binary label available at training.** Images from the same set come from the same domain.

array of biomedical datasets spanning diverse domains and segmentation protocols. In most realistic scenarios, only a subset of the labels (often one label) in a protocol are available in each dataset, and most often only one protocol. At each training iteration, we first sample a dataset from the biomedical collection. If a dataset contains multiple protocols (which is rarely available in public data), we sample a protocol, and then sample a specific label task t within that protocol. Finally, we randomly sample a *set* of images from that dataset, with that label segmented.

At each iteration, for a set of images $\{x_s\}$ with associated ground-truth binary labels $\{y_s\}$, we sample M protocols containing at most K labels each, and predict label maps $\{\hat{y}_{m,s}\}$ for *all* M protocols.

Loss function. We design a loss function that *encourages the label maps of any produced protocol to be consistent across the image set*. The loss function also enables learning from data with only a subset of labels segmented in each protocol, and encourages diversity of predicted candidate protocols. We develop this further in the Supplementary Section B.

We define $d_{m,k}(\{\hat{y}_{s,m,k}\}, \{y_s\}) = \mathbb{E}_s[\mathcal{L}_{Dice}(\hat{y}_{s,m,k}, y_s)]$ as the average Dice score for a protocol m and label k across the samples s, where $\hat{y}_{m,s,k}$ is the binary map of label k of prediction $\hat{y}_{m,s}$. Denoting $\mathcal{T}$ as the set of all possible tasks and $\mathcal{S}$ the set of all image sets, we optimize model parameters θ_f and θ_h by minimizing the loss function

$$\mathcal{L}(\theta_f, \theta_h; \mathcal{T}) = \mathbb{E}_{\mathcal{T}}\mathbb{E}_{\mathcal{S}}[\mathcal{L}_{seg}(\{\hat{y}_{s,m,k}\}, y_s^t)], \tag{5}$$

$$\text{with} \quad \mathcal{L}_{seg}(\{\hat{y}_{s,m,k}\}, y_s^t) = \min_{m,k} d_{m,k}(\{\hat{y}_{s,m,k}\}, \{y_s^t\}). \tag{6}$$

By only penalizing the segmentation of the best performing predicted protocol m and label k, the loss function encourages diverse label map samples $\hat{y}_m$ – *at least one* candidate segmentation map matches the ground-truth binary label [58, 96]. By averaging the loss terms across the set, the loss encourages label k of protocol m to refer to the same region in each image. Figure 4 shows examples of the inputs and their corresponding binary labels used in the loss.

Augmentations. We apply standard data augmentations to improve generalization, such as Gaussian noise, blur, contrast changes, affine, and elastic transforms. Building on recent strategies [16], we distinguish between two types of augmentations, *in-task* augmentations and *task* augmentations. The *in-task* augmentations are applied independently to each element in a set and aim to increase the diversity of sets. The *task* augmentations are applied consistently across elements of the same set and aid in increasing the number of protocols available at training. A complete list of augmentations and parameters is provided in the supplemental material in F.3.

Synthetic Data. To improve the generalization to new domains, we use synthetic data [13, 16, 43] build on *Anatomix* [27]. First, we create maps by sampling binary segmentations from TotalSegmentator [118]. To simulate images from the same set representing the same biomedical region, we first sample a single label map, which is shared across the elements of one set. We then create label maps by applying affine and elastic transforms to the original label map independently. We assign an intensity to all pixels in a given label, and apply several related augmentations to create a synthetic corresponding image set.

3.2 Implementation details

For function $f_{\theta_f}(\cdot)$, we use a UNet-like architecture, with convolutional layers of 32 features followed by *PReLU* activation [39]. The function $h_{\theta_h}(\cdot)$ is a series of convolution layers with skip connections. We use the SoftMax function across the K dimension to obtain multi-label segmentation maps, with non-overlapping labels. This ensures that our protocols are *complete*, assigning a label to each image pixel, instead of *partial*, assigning labels to only specific regions of the image. During training, we sample the maximum number of labels K, the number of protocols M, and the set size S uniformly from a fixed range: $K \in [5, 40], M \in [5, 15], S \in [2, 5]$. Within the same batch, different sets come from different domains. At inference time, K and M can be chosen by the user, while S is determined by the number of images in the set to segment. We use the AdamW optimizer [74] with a learning rate of 0.0001 [56].

4 Experiments

We evaluate Pancakes on a broad battery of biomedical segmentation tasks, with three goals: (1) verify that each *protocol* generated by the model is semantically consistent across images; (2) quantify how well Pancakes matches manual segmentations of a provided protocol; and (3) study how design choices during training and inference affect diversity and accuracy. We pay special attention to datasets and imaging domains unseen during training.

Evaluation. The evaluation of a method's ability to produce segmentation maps for different plausible protocols, consistently across subjects, is substantially more challenging than evaluation in standard (fixed protocol) segmentation tasks. In most test datasets, only one protocol is provided and only one label is available from that protocol.

We start by assessing if any of the predicted protocols produce a label that closely matches a ground-truth label. We then capture whether this label is consistent across the set. Specifically, we compute the Dice score [107, 109] between the available ground-truth y_s^t and the labels produced for each label map and each image in the set $\hat{y}_{s,m,k}$. We then compute the average across subjects in the set: $\mathbb{E}_s\{Dice(\hat{y}_{s,m,k}, y_s^t)\}]$. Finally, we record the produced label that performs best, identified by a specific map m and label k. We call this *Set Dice*:

$$\max_{m,k} \mathbb{E}_s\{Dice(\{\hat{y}_{s,m,k}\}, y_s^t)\}. \tag{7}$$

This metric captures: (1) Consistency: a given label should represent the same region in all images in a set; and (2) Accuracy: for each individual image, how well does the best label match the ground-truth segmentation?

Because our goal is to generate a family of anatomically coherent multi-protocol label maps, no single user-provided mask can serve as a universal "ground-truth". Therefore, overlap metrics such as Dice score offer only a partial assessment. We complement our quantitative evaluation with visualizations that illustrate the range of protocols generated by Pancakes. We assess these examples in two aspects: (1) each protocol should be consistent across subjects; and (2) different protocols should generate different, yet still anatomically plausible, partitions. Inevitably, some protocols will resonate more with readers' expectations than others. Our goal in this visualization is to demonstrate the diversity and semantic coherence Pancakes can achieve.

Data. We train Pancakes on a large, diverse collection of biomedical data, and evaluate the multi-protocol segmentations produced on images from held-out datasets. We use Megamedical [16, 96, 123], which covers many biomedical domains [2, 3, 6, 11, 12, 14, 16, 19, 22, 29, 31, 32, 35, 37, 41, 42, 47, 49, 51, 52, 54, 61–67, 69, 71–73, 75, 78–82, 87, 93, 94, 98, 101, 103, 105, 106, 108, 110, 127, 129, 132, 133].This dataset has diverse anatomies such as thoracic organs [75], brain [80], eye [78]; and diverse modalities including XRay [47], CT [75], ultrasound [66] and fundus images [46]. We partition this collection into three subgroups. *Training datasets* are used at training time, including model weight optimization and backpropagation. A complete list of the training datasets can be found Table 4 in the supplemental material. *Development datasets*, ACDC [11], PanDental [2], and SpineWeb [132], are *not* used in training but are to evaluate models during development. *Held-out datasets*. These datasets are only used for final evaluation. They include QUBIQ Prostate dataset [83], SCD [94], WBC [133], BUID [3], LIDC-IDRI [5], DDTI [90], and STARE [46].

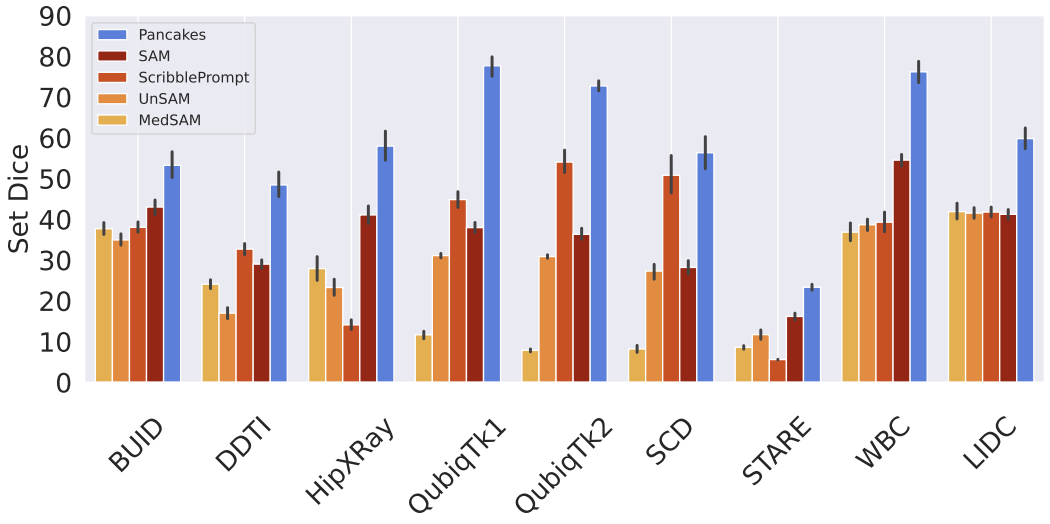


Figure 5: **Pancakes generalizes well to unseen datasets.** We evaluate Pancakes and baselines with the *Set Dice* metric.

The images within each dataset are also split into *training*, *validation*, and *test* splits. We train the models on the *training* split of the *training* datasets. We used the *training* split of the *development* datasets to monitor out-of-distribution capabilities. We report results on the *test* splits of both the *development* (in the supplemental material) and *held-out* datasets. We split the dataset based on subjects, and ensured that there was no train/validation/test subject cross-contamination.

We also synthesized $120,000$ training image-segmentation pairs as described in Section 3. We refer to these examples as *Anatomix* data.

Benchmarks. To our knowledge, there are no methods that attempt the same task as Pancakes. We compare our work to four methods that can be used to produce segmentation maps for new images, but none of these were designed to produce *multiple* protocols *consistently* across subjects.

SAM [58, 97]: the Segment Anything Model (SAM) is an interactive image segmentation model trained mostly on natural images. SAM involves a whole-image-segmentation mode, which produces a grid of simulated clicks over the whole image, removes high-overlapping regions and selects the most likely masks using a confidence threshold. We use this mode, and produce diverse whole-image segmentations by varying the confidence threshold.

ScribblePrompt [123]: ScribblePrompt (SP), is an interactive segmentation tool trained on Megamedical. We use the SP-SAM, which performs best on clicks, and obtain multi-protocol and multi-label using the same whole-image strategy we used for SAM.

MedSAM [76]: trained specifically on medical data, this model uses a SAM-like architecture and is optimized for bounding box interactions. We produce whole-image segmentations using the SAM whole-image segmentation mode.

UnSAM [116]: trained on a curated set of natural images, this model produces segmentation masks using a DINO [20] backbone (pretrained ResNet50 [40]) and the Mask2Former [21] mask decoder. UnSAM produces a list of labels to serve as label maps. We use the *UnSAM+* model version, trained on a part of the SA-1B dataset [58]. To produce segmentations with diverse protocols and labels, we use the UnSAM's existing whole-image segmentation scheme.

5 Results

For our main evaluation, all error bars are 95% confidence interval using 1000 bootstraps of all sets. Additional per-dataset performance and ablations are shown in the supplemental material.

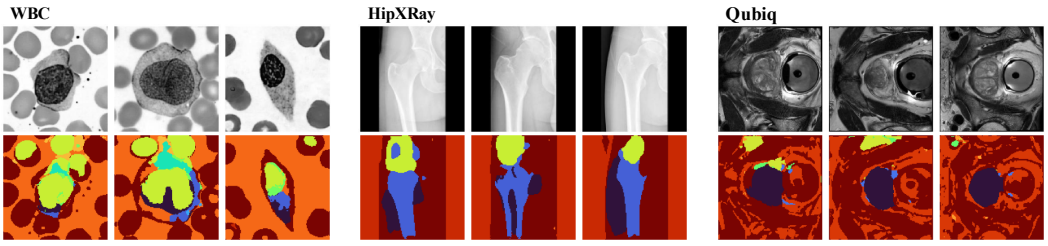


Figure 6: **Pancakes set consistency.** For the same protocol, each label, identified by color, is assigned to similar structures across a set of images.

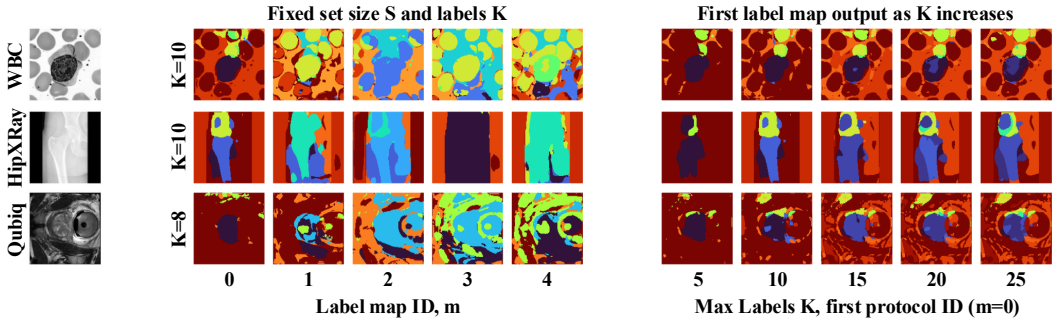


Figure 7: **Effect of M and K.** Left: Input image for the first subject in the set ($s = 0$) with set size $S = 3$. Middle: Pancakes can produce diverse label maps with fixed K labels per protocol. Right: Increasing the maximum number of labels K tends to lead to finer structures in the produced label maps.

Quantitatively, Figure 5 shows that Pancakes outperforms the baselines on all unseen datasets, often by a margin of more than 20 Dice points. For a more detailed assessment, we separately report accuracy, consistency, and the effects of hyperparameters, and visualize results in several scenarios.

Accuracy versus consistency. Figure 8 shows that Pancakes performs well on *both* segmentation accuracy and semantic consistency across images. For individual accuracy (set size $S = 1$), which does not penalize semantic consistency across images, Pancakes performs similarly to SAM, and is superior to all the other baselines. As the set size S increases, all baselines fail to produce semantically *consistent* segmentations, while Pancakes segmentations remain consistent across the set, leading

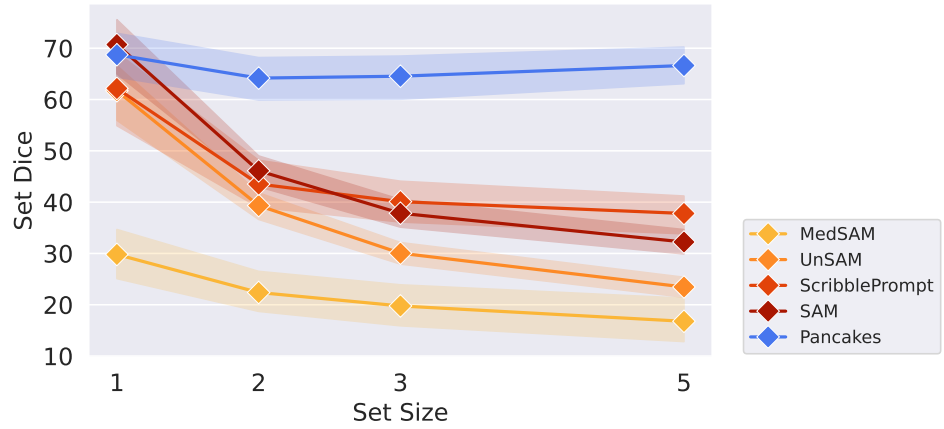


Figure 8: **Pancakes is both consistent and accurate compared to baselines.** When evaluated solely on accuracy ($S = 1$), Pancakes is comparable to SAM and outperforms baselines. As S increases, Pancakes is the only model whose performance is not degraded.

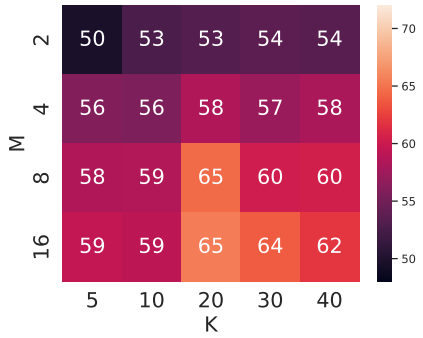


Figure 9: **Influence of segmentation maps M and labels K on segmentation quality for $S = 3$.**

Table 1: **Inference time and number of parameters.** Pancakes has substantially fewer parameters and is significantly faster than baseline methods.

	#Param	S=1 (sec.)	S=3 (sec.)
Ours	**0.22 M**	**0.10 ± 0.04**	**0.12 ± 0.03**
SAM	641M	3.13±0.16	2.94 ± 0.24
SP	93.7M	1.99 ± 0.13	1.85 ± 0.18
MedSAM	93.7M	2.12 ± 0.17	1.94 ± 0.18
UnSAM	23M	0.54 ± 0.012	0.45 ± 0.03

to a steady *Set Dice*. We hypothesize that SAM is better than biomedical baselines because it was trained on a wider variety of labels and images. Therefore, SAM is less prone to task-overfitting. In the Supplementary Section C, we also report results for various M and K, and also evaluate using the *Set Surface Distance* and *Set IoU* metrics.

Figure 6 illustrates Pancakes predictions for a fixed protocol, and highlights visually the semantic consistency across the images of a set. Additional visualizations with the baselines are shown in Section C.1.

Overall, Pancakes outperforms or matches the other methods in producing plausible protocols, and outperforms all methods by a substantial margin in producing semantically consistent protocols.

Influence of M and K. Figure 7 illustrates the diversity of protocols captured in Pancakes segmentation outputs. They are produced by fixing the set size S and the maximum number of labels K and performing one forward pass through the network. The label maps are different from one another across protocol ID m and maximum labels K, capturing structures at various granularity levels. Figure 7 also shows that increasing the number of labels K per protocol results in segmentations of finer structures.

Figure 9 captures the quantitative effects of the number of label maps M and maximum label number K for a fixed set size S. Producing more protocols leads to better performance in general, while performance as a function of K is more variable. We hypothesize that this arises from the non-overlapping nature of labels within each protocol. Ambiguous regions that could belong to multiple labels require separate protocols to represent each plausible interpretation, so having more protocols enables the model to better capture such ambiguities.

Efficiency. We study runtime requirements by running all the models on the HipXRay [37] test split. Table 1 shows the average run time across 1000 trials. Pancakes is substantially faster than all baselines, because of its efficient fully-convolutional architecture, leading to a leaner model and fewer parameters.

Analysis. We use *Set Dice* with $S = 3$, and use $M = 8$ and $K = 20$ for Pancakes. As supplementary Figure 15 shows, performance on the development datasets saturates at $M = 8$ and $K = 20$, with only marginal gains achieved by further raising M. We therefore chose to use these parameters as a balance between performance and the number of structures clinical users might actually expect in practice.

Influence of synthetic data. We compare versions of Pancakes trained with real data only, synthetic data only, or both. When trained with both real and synthetic data, Pancakes consistently outperforms the other variants in Dice score ($p < 0.05$ with a paired Student-t test), for $M = 16$, as shown in Table 2. We find that for $M = 8$, the performance change is not statistically significant.

Pancakes and interactive segmentation. If a user has a particular segmentation protocol in mind, but there are no existing tools for it, they can choose the protocol from Pancakes' outputs that best

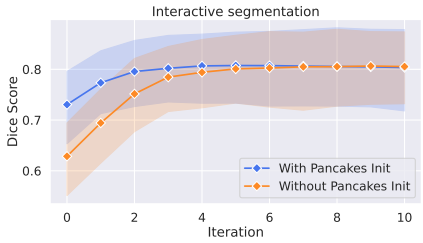


Figure 10: **Interactive Segmentation with Pancakes initialization.** Pancakes can be used as an initialization for interactive segmentation to obtain high quality labels faster.

Table 2: **Influence of synthetic data across set size.** For $M = 16$, training with both synthetic data and Megamedical yields a small yet consistent improvement over only training with Megamedical.

	Both	Megamedical	Synthetic
S = 1	**73.2 ± 5.5**	71.1 ± 6.2	56.3 ± 4.9
S = 2	**67.3 ± 6.0**	65.8 ± 6.7	45.8 ± 5.0
S = 3	**67.4 ± 6.0**	65.7 ± 6.8	44.3 ± 5.4
S = 5	**68.4 ± 5.6**	67.4 ± 6.9	42.7 ± 5.5

aligns with their intended use. This choice will often suffice. In cases where Pancakes' segmentations are not sufficiently accurate–for example, when targets are far from the training distribution–the segmentation maps can offer excellent initializations to interactive segmentation systems such as ScribblePrompt [123]. To demonstrate this, we compared using ScribblePrompt with and without initialization using Pancakes' predictions: (1) on average, Pancakes' predictions can be improved by 5 Dice points with a *single interactive click*, for set sizes larger than one. (2) on average, using ScribblePrompt with a Pancakes-initialized segmentation reduces the number of required interactions by half. If used alone, it takes ScribblePrompt 5 to 8 clicks for the prediction quality to plateau, while, when initialized with Pancakes' predictions, ScribblePrompt can reach the same results in 3 to 4 clicks.

6 Assumptions & Social Impact

In this work, we made several core assumptions. First, we assume that the user has an idea of the desired labels and can select a few K values *a priori*. Pancakes is intended for biomedical experts. We assume that as they use the tool, they can identify the mapping between a label and a known structure. Second, we assume that visualizing several protocols simultaneously is reasonable. Third, we assume that while we trained and evaluated on a diverse set of medical images, we likely did not capture all biomedical image types and domains that a user might encounter. Fourth, Pancakes is not intended to replace existing clinically validated segmentation protocols. If there is an existing tool for a particular protocol, we would advise using it.

We aim for this work to inspire a new approach to using foundation models in biomedical imaging. Pancakes is designed to be efficient and accessible, especially in resource-constrained settings. It can support both data annotation and exploratory analysis, as well as serve as a tool for developing future foundation models. However, this version is intended for research use only. While trained on a broad collection of small biomedical datasets, we have not evaluated it for potential societal biases.

7 Conclusion

We introduced Pancakes, a new framework that predicts segmentation maps for multiple protocols in previously unseen biomedical imaging domains, with each protocol being semantically consistent across images. Pancakes estimates a distribution over plausible label map protocols and provides a mechanism to sample multiple segmentation maps from this distribution.

Predicting plausible segmentations in a new domain, without prior knowledge of the number of labels or their associated shapes, is a challenging task. Our experiments demonstrate that Pancakes achieves state-of-the-art performance on seven held-out datasets. We believe this work addresses an important, previously unaddressed problem, enabling biomedical researchers to segment entire collections of images without requiring manual annotations or example segmentations. This can, in turn, substantially speed up downstream biomedical studies.

8 Acknowledgement

We would like to thank Neel Dey for the very helpful discussions and feedback. This research was supported 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. Some of the computation resources required for this research was performed on computational hardware generously provided by the Massachusetts Life Sciences Center.

References

[1] Topbrain 2025 grand challenge. `https://topbrain2025.grand-challenge.org`. Accessed: 2025-08-23.

[2] A. H. Abdi, S. Kasaei, and M. Mehdizadeh. Automatic segmentation of mandible in panoramic x-ray. *Journal of Medical Imaging*, 2(4):044003, 2015.

[3] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy. Dataset of breast ultrasound images. *Data in Brief*, 28:104863, 2020.

[4] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

[5] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, 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] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

[7] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.

[8] I. Balazevic, D. Steiner, N. Parthasarathy, R. Arandjelović, and O. Henaff. Towards in-context scene understanding. *Advances in Neural Information Processing Systems*, 36, 2024.

[9] S. Bano, F. Vasconcelos, L. M. Shepherd, E. Vander Poorten, T. Vercauteren, S. Ourselin, A. L. David, J. Deprest, and D. 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.

[10] C. F. Baumgartner, K. C. Tezcan, K. Chaitanya, A. M. Hötker, U. J. Muehlematter, K. Schawkat, A. S. Becker, O. Donati, and E. Konukoglu. Phiseg: Capturing uncertainty in medical image segmentation. In *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.

[11] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. 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.

[12] P. Bilic, P. F. Christ, E. Vorontsov, G. Chlebus, H. Chen, Q. Dou, C.-W. Fu, X. Han, P.-A. Heng, J. Hesser, et al. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019.

[13] B. Billot, D. N. Greve, O. Puonti, A. Thielscher, K. Van Leemput, B. Fischl, A. V. Dalca, J. E. Iglesias, et al. Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining. *Medical image analysis*, 86:102789, 2023.

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

[15] M. Buda, A. Saha, and M. 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.

[16] V. I. Butoi, J. J. G. Ortiz, T. Ma, M. R. Sabuncu, J. Guttag, and A. V. Dalca. Universeg: Universal medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21438–21451, 2023.

[17] J. C. Caicedo, A. Goodman, K. W. Karhohs, B. A. Cimini, J. Ackerman, M. Haghighi, C. Heng, T. Becker, M. Doan, C. McQuin, M. Rohban, S. Singh, and A. E. Carpenter. Nucleus segmentation across imaging experiments: the 2018 Data Science Bowl. *Nature Methods*, 16(12):1247–1253, Dec. 2019.

[18] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[19] A. Cardon, S. Saalfeld, S. Preibisch, B. Schmid, A. Cheng, J. Pulokas, P. Tomancak, and V. Hartenstein. Isbi challenge: Segmentation of neuronal structures in em stacks.

[20] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[21] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[22] N. C. F. Codella, D. A. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. K. Mishra, H. Kittler, and A. 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.

[23] S. Czolbe, K. Arnavaz, O. Krause, and A. 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.

[24] D. de Geus, P. Meletis, C. Lu, X. Wen, and G. Dubbelman. Part-aware panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5485–5494, 2021.

[25] E. Decenciere, G. Cazuguel, X. Zhang, G. Thibault, J.-C. Klein, F. Meyer, B. Marcotegui, G. Quellec, M. Lamard, R. Danno, et al. Teleophta: Machine learning and image processing methods for teleophthalmology. *Irbm*, 34(2):196–203, 2013.

[26] A. Degerli, M. Zabihi, S. Kiranyaz, T. Hamid, R. Mazhar, R. Hamila, and M. Gabbouj. Early detection of myocardial infarction in low-quality echocardiography. *IEEE Access*, 9:34442–34453, 2021.

[27] N. Dey, B. Billot, H. E. Wong, C. J. Wang, M. Ren, P. E. Grant, A. V. Dalca, and P. Golland. Learning general-purpose biomedical volume representations using randomized synthesis, 2024.

[28] B. Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

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

[30] Y. Gao, D. Liu, Z. Li, Y. Li, D. Chen, M. Zhou, and D. N. Metaxas. Show and segment: Universal medical image segmentation via in-context learning. *arXiv preprint arXiv:2503.19359*, 2025.

[31] S. Gerhard, J. Funke, J. Martel, A. Cardona, and R. Fetter. Segmented anisotropic ssTEM dataset of neural tissue. *figshare*, pages 0–0, 11 2013.

[32] R. L. Gollub, J. M. Shoemaker, M. D. King, T. White, S. Ehrlich, S. R. Sponheim, V. P. Clark, J. A. Turner, B. A. Mueller, V. 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.

[33] I. S. Gousias, A. D. Edwards, M. A. Rutherford, S. J. Counsell, J. V. Hajnal, D. Rueckert, and A. 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.

[34] I. S. Gousias, D. Rueckert, R. A. Heckemann, L. E. Dyet, J. P. Boardman, A. D. Edwards, and A. Hammers. Automatic segmentation of brain mris of 2-year-olds into 83 regions of interest. *Neuroimage*, 40(2):672–684, 2008.

[35] S. Graham, Q. D. Vu, S. E. A. Raza, A. Azam, Y. W. Tsang, J. T. Kwak, and N. Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019.

[36] D. N. Greve, B. Billot, D. Cordero, A. Hoopes, M. Hoffmann, A. V. Dalca, B. Fischl, J. E. Iglesias, and J. C. Augustinack. A deep learning toolbox for automatic segmentation of subcortical limbic structures from mri images. *Neuroimage*, 244:118610, 2021.

[37] D. Gut. X-ray images of the hip joints. 1, July 2021. Publisher: Mendeley Data.

[38] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI brainlesion workshop*, pages 272–284. Springer, 2021.

[39] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[40] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[41] N. Heller, F. Isensee, K. H. Maier-Hein, X. Hou, C. Xie, F. Li, Y. Nan, G. Mu, Z. Lin, M. 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.

[42] M. R. Hernandez Petzsche, E. de la Rosa, U. Hanning, R. Wiest, W. Valenzuela, M. Reyes, M. Meyer, S.-L. Liew, F. Kofler, I. Ezhov, et al. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific data*, 9(1):762, 2022.

[43] M. Hoffmann, B. Billot, D. N. Greve, J. E. Iglesias, B. Fischl, and A. V. Dalca. Synthmorph: learning contrast-invariant registration without acquired images. *IEEE transactions on medical imaging*, 41(3):543–558, 2021.

[44] A. Hoopes, V. I. Butoi, J. V. Guttag, and A. V. Dalca. Voxelprompt: A vision-language agent for grounded medical image analysis. *arXiv preprint arXiv:2410.08397*, 2024.

[45] A. Hoopes, M. Hoffmann, D. N. Greve, B. Fischl, J. Guttag, and A. V. Dalca. Learning the effect of registration hyperparameters with hypermorph. volume 1, pages 1–30, 2022.

[46] A. Hoover, V. Kouznetsova, and M. 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.

[47] H. in the Loop. Teeth segmentation dataset.

[48] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

[49] Y. Ji, H. Bai, J. Yang, C. Ge, Y. Zhu, R. Zhang, Z. Li, L. Zhang, W. Ma, X. Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *arXiv preprint arXiv:2206.08023*, 2022.

[50] Q. Jin, Z. Meng, T. D. Pham, Q. Chen, L. Wei, and R. Su. Dunet: A deformable network for retinal vessel segmentation. *Knowledge-Based Systems*, 178:149–162, 2019.

[51] R. Karim, R. J. Housden, M. Balasubramaniam, Z. Chen, D. Perry, A. Uddin, Y. Al-Beyatti, E. Palkhi, P. Acheampong, S. 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.

[52] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan, B. Baydar, D. Lachinov, S. Han, J. Pauli, F. Isensee, M. Perkonigg, R. Sathish, R. Rajan, D. Sheet, G. Dovletov, O. Speck, A. Nürnberger, K. H. Maier-Hein, G. Bozdağı Akar, G. Ünal, O. Dicle, and M. A. Selver. CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, Apr. 2021.

[53] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan, B. Baydar, D. Lachinov, S. Han, J. Pauli, F. Isensee, M. Perkonigg, R. Sathish, R. Rajan, D. Sheet, G. Dovletov, O. Speck, A. Nürnberger, K. H. Maier-Hein, G. Bozdağı Akar, G. Ünal, O. Dicle, and M. A. Selver. CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021.

[54] A. E. Kavur, M. A. Selver, O. Dicle, M. Barış, and N. S. Gezer. CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data. Apr. 2019.

[55] A. E. Kavur, M. A. Selver, O. Dicle, M. Barış, and N. S. Gezer. CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data, Apr. 2019.

[56] D. P. Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[57] S. Kiranyaz, A. Degerli, T. Hamid, R. Mazhar, R. E. F. Ahmed, R. Abouhasera, M. Zabihi, J. Malik, R. Hamila, and M. Gabbouj. Left ventricular wall motion estimation by active polynomials for acute myocardial infarction detection. *IEEE Access*, 8:210301–210317, 2020.

[58] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[59] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. Eslami, D. Jimenez Rezende, and O. Ronneberger. A probabilistic u-net for segmentation of ambiguous images. *Advances in neural information processing systems*, 31, 2018.

[60] S. A. Kohl, B. Romera-Paredes, K. H. Maier-Hein, D. J. Rezende, S. Eslami, P. Kohli, A. Zisserman, and O. Ronneberger. A hierarchical probabilistic u-net for modeling multi-scale ambiguities. *arXiv preprint arXiv:1905.13077*, 2019.

[61] M. Krönke, C. Eilers, D. Dimova, M. Köhler, G. Buschner, L. Schweiger, L. Konstantinidou, M. Makowski, J. Nagarajah, N. 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.

[62] H. J. Kuijf, J. M. Biesbroek, J. De Bresser, R. Heinen, S. Andermatt, M. Bento, M. Berseth, M. Belyaev, M. J. Cardoso, A. 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.

[63] M. Kuklisova-Murgasova, P. Aljabar, L. Srinivasan, S. J. Counsell, V. Doria, A. Serag, I. S. Gousias, J. P. Boardman, M. A. Rutherford, A. D. Edwards, et al. A dynamic 4d probabilistic atlas of the developing brain. *NeuroImage*, 54(4):2750–2763, 2011.

[64] Z. Lambert, C. Petitjean, B. Dubray, and S. 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.

[65] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, page 12, 2015.

[66] S. Leclerc, E. Smistad, J. Pedrosa, A. Østvik, F. Cervenansky, F. Espinosa, T. Espeland, E. A. R. Berg, P.-M. Jodoin, T. 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.

[67] G. Lemaître, R. Martí, J. Freixenet, J. C. Vilanova, P. M. Walker, and F. Meriaudeau. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review. *Computers in biology and medicine*, 60:8–31, 2015.

[68] F. Li, H. Zhang, P. Sun, X. Zou, S. Liu, J. Yang, C. Li, L. Zhang, and J. Gao. Semantic-sam: Segment and recognize anything at any granularity. *arXiv preprint arXiv:2307.04767*, 2023.

[69] M. Li, Y. Zhang, Z. Ji, K. Xie, S. Yuan, Q. Liu, and Q. Chen. Ipn-v2 and octa-500: Methodology and dataset for retinal image segmentation. *arXiv preprint arXiv:2012.07261*, 2020.

[70] P. Liskowski and K. Krawiec. Segmenting retinal blood vessels with deep neural networks. *IEEE transactions on medical imaging*, 35(11):2369–2380, 2016.

[71] G. Litjens, R. Toth, W. van de Ven, C. Hoeks, S. Kerkstra, B. van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014.

[72] V. Ljosa, K. L. Sokolnicki, and A. E. Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature methods*, 9(7):637–637, 2012.

[73] M. T. Löffler, A. Sekuboyina, A. Jacob, A.-L. Grau, A. Scharr, M. El Husseini, M. Kallweit, C. Zimmer, T. Baum, and J. S. Kirschke. A vertebral segmentation dataset with fracture grading. *Radiology: Artificial Intelligence*, 2(4):e190138, 2020.

[74] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[75] X. Luo, W. Liao, J. Xiao, T. Song, X. Zhang, K. Li, G. Wang, and S. Zhang. Word: Revisiting organs segmentation in the whole abdominal region. *arXiv preprint arXiv:2111.02403*, 2021.

[76] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[77] Q. Ma, L. Li, E. C. Robinson, B. Kainz, and D. Rueckert. Weakly supervised learning of cortical surface reconstruction from segmentations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 766–777. Springer, 2024.

[78] Y. Ma, H. Hao, J. Xie, H. Fu, J. Zhang, J. Yang, Z. Wang, J. Liu, Y. Zheng, and Y. Zhao. Rose: a retinal oct-angiography vessel segmentation dataset and new model. *IEEE Transactions on Medical Imaging*, 40(3):928–939, 2021.

[79] J. A. Macdonald, Z. Zhu, B. Konkel, M. Mazurowski, W. Wiggins, and M. Bashir. Duke liver dataset (MRI) v2, Apr. 2023.

[80] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. 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.

[81] K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury, et al. The parkinson progression marker initiative (ppmi). *Progress in neurobiology*, 95(4):629–635, 2011.

[82] M. A. Mazurowski, K. Clark, N. M. Czarnek, P. Shamsesfandabadi, K. B. Peters, and A. 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.

[83] B. Menze, L. Joskowicz, S. Bakas, A. Jakab, E. Konukoglu, A. Becker, A. Simpson, and R. D. Quantification of uncertainties in biomedical image quantification 2021. *4th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2021)*, 2021.

[84] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[85] R. Merdietio Boedi, S. Shepherd, F. Oscandar, S. Mânica, and A. Franco. 3d segmentation of dental crown for volumetric age estimation with cbct imaging. *International Journal of Legal Medicine*, 137(1):123–130, 2023.

[86] M. Monteiro, L. Le Folgoc, D. Coelho de Castro, N. Pawlowski, B. Marques, K. Kamnitsas, M. van der Wilk, and B. Glocker. Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *Advances in Neural Information Processing Systems*, 33:12756–12767, 2020.

[87] A. Montoya, Hasnin, kaggle446, shirzad, W. Cukierski, and yffud. Ultrasound nerve segmentation, 2016.

[88] J. Myers-Dean, J. Reynolds, B. Price, Y. Fan, and D. Gurari. Spin: Hierarchical segmentation with subpart granularity in natural images. *arXiv preprint arXiv:2407.09686*, 2024.

[89] K. Payette, P. de Dumast, H. Kebiri, I. Ezhov, J. C. Paetzold, S. Shit, A. Iqbal, R. Khan, R. Kottke, P. 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.

[90] L. Pedraza, C. Vargas, F. Narváez, O. Durán, E. Muñoz, and E. Romero. An open access thyroid ultrasound image database. In *10th International symposium on medical information processing and analysis*, volume 9287, pages 188–193. SPIE, 2015.

[91] L. Pedraza, C. Vargas, F. Narváez, O. Durán, E. Muñoz, and E. Romero. An open access thyroid ultrasound image database. In *10th International symposium on medical information processing and analysis*, volume 9287, pages 188–193. SPIE, 2015.

[92] H. Peiris, M. Hayat, Z. Chen, G. Egan, and M. Harandi. A robust volumetric transformer for accurate 3d tumor segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 162–172. Springer, 2022.

[93] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Meriaudeau. Indian diabetic retinopathy image dataset (idrid), 2018.

[94] P. Radau, Y. Lu, K. Connelly, G. Paul, A. Dick, and G. Wright. Evaluation framework for algorithms segmenting short axis cardiac mri. *The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge*, 49, 2009.

[95] A. Rahman, J. M. J. Valanarasu, I. Hacihaliloglu, and V. 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.

[96] M. Rakic, H. E. Wong, J. J. G. Ortiz, B. A. Cimini, J. V. Guttag, and A. V. Dalca. Tyche: Stochastic in-context learning for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11159–11173, 2024.

[97] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[98] B. Rister, D. Yi, K. Shivakumar, T. Nobashi, and D. L. Rubin. CT-ORG, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7(1):381, Nov. 2020.

[99] O. Ronneberger, P. Fischer, and T. 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.

[100] F. Rossant, M. Badellino, A. Chavillon, I. Bloch, and M. Paques. A morphological approach for vessel segmentation in eye fundus images, with quantitative evaluation. *Journal of Medical Imaging and Health Informatics*, 1(1):42–49, 2011.

[101] A. Saporta, X. Gui, A. Agrawal, A. Pareek, S. Truong, C. Nguyen, V.-D. Ngo, J. Seekins, F. G. Blankenberg, A. Ng, et al. Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation. *MedRxiv*, 2021.

[102] A. Schmidt, P. Morales-Álvarez, and R. 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.

[103] C. Seibold, S. Reiß, S. Sarfraz, M. A. Fink, V. Mayer, J. Sellner, M. S. Kim, K. H. Maier-Hein, J. Kleesiek, and R. Stiefelhagen. Detailed annotations of chest x-rays via ct projection for report understanding. In *Proceedings of the 33th British Machine Vision Conference (BMVC)*, 2022.

[104] A. Serag, P. Aljabar, G. Ball, S. J. Counsell, J. P. Boardman, M. A. Rutherford, A. D. Edwards, J. V. Hajnal, and D. Rueckert. Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression. *Neuroimage*, 59(3):2255–2265, 2012.

[105] A. A. A. Setio, A. Traverso, T. De Bel, M. S. Berens, C. Van Den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. 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.

[106] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

[107] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[108] Y. Song, J. Zheng, L. Lei, Z. Ni, B. Zhao, and Y. Hu. CT2US: Cross-modal transfer learning for kidney segmentation in ultrasound images with synthesized data. *Ultrasonics*, 122:106706, 2022.

[109] T. Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biol Skrifter/Kongelige Danske Videnskabernes Selskab.*, 5:1, 1948.

[110] J. Staal, M. D. Abràmoff, M. Niemeijer, M. A. Viergever, and B. Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4):501–509, 2004.

[111] R. Tanno, A. Saeedi, S. Sankaranarayanan, D. C. Alexander, and N. 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.

[112] P. V. Tran. A fully convolutional neural network for cardiac segmentation in short-axis mri. *arXiv preprint arXiv:1604.00494*, 2016.

[113] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[114] X. Wang, S. Li, K. Kallidromitis, Y. Kato, K. Kozuka, and T. Darrell. Hierarchical open-vocabulary universal image segmentation. *Advances in Neural Information Processing Systems*, 36, 2024.

[115] X. Wang, W. Wang, Y. Cao, C. Shen, and T. 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.

[116] X. Wang, J. Yang, and T. Darrell. Segment anything without supervision. *arXiv preprint arXiv:2406.20081*, 2024.

[117] X. Wang, X. Zhang, Y. Cao, W. Wang, C. Shen, and T. Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023.

[118] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. T. Boll, J. Cyriac, S. Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5), 2023.

[119] W. Wells III, W. Grimson, R. Kikinis, and F. Jolesz. Adaptive segmentation of mri data. *IEEE Transactions on Medical Imaging*, 15:429–442, 1996.

[120] J. Wolleb, F. Bieder, R. Sandkühler, and P. C. Cattin. Diffusion models for medical anomaly detection. In *International Conference on Medical image computing and computer-assisted intervention*, pages 35–45. Springer, 2022.

[121] J. Wolleb, R. Sandkühler, F. Bieder, P. Valmaggia, and P. C. Cattin. Diffusion Models for Implicit Image Segmentation Ensembles. In *Medical Imaging with Deep Learning*, 2021.

[122] H. E. Wong, J. J. G. Ortiz, J. Guttag, and A. V. Dalca. Multiverseg: Scalable interactive segmentation of biomedical imaging datasets with in-context guidance. *arXiv preprint arXiv:2412.15058*, 2024.

[123] H. E. Wong, M. Rakic, J. Guttag, and A. V. Dalca. Scribbleprompt: Fast and flexible interactive segmentation for any medical image. *arXiv preprint arXiv:2312.07381*, 2023.

[124] J. Wu and M. Xu. One-prompt to segment all medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11302–11312, 2024.

[125] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018.

[126] L. Zbinden, L. Doorenbos, T. Pissas, A. T. Huber, R. Sznitman, and P. Márquez-Neila. Stochastic segmentation with conditional categorical diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1119–1129, 2023.

[127] Y. Zhang, M. Xian, H.-D. Cheng, B. Shareef, J. Ding, F. Xu, K. Huang, B. Zhang, C. Ning, and Y. Wang. Busis: A benchmark for breast ultrasound image segmentation. In *Healthcare*, volume 10, page 729. MDPI, 2022.

[128] Y. Zhang, M. Xian, H.-D. Cheng, B. Shareef, J. Ding, F. Xu, K. Huang, B. Zhang, C. Ning, and Y. Wang. Busis: A benchmark for breast ultrasound image segmentation. In *Healthcare*, volume 10, page 729. MDPI, 2022.

[129] Q. Zhao, S. Lyu, W. Bai, L. Cai, B. Liu, M. Wu, X. Sang, M. Yang, and L. Chen. A multimodality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation. *CoRR*, abs/2207.06799, 2022.

[130] T. Zhao, Y. Gu, J. Yang, N. Usuyama, H. H. Lee, T. Naumann, J. Gao, A. Crabtree, B. Piening, C. Bifulco, et al. Biomedparse: a biomedical foundation model for image parsing of everything everywhere all at once. *arXiv preprint arXiv:2405.12971*, 2024.

[131] Z. Zhao, Y. Zhang, C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie. One model to rule them all: Towards universal segmentation for medical images with text prompts. *arXiv preprint arXiv:2312.17183*, 2023.

[132] G. Zheng, C. Chu, D. L. Belavỳ, B. Ibragimov, R. Korez, T. Vrtovec, H. Hutt, R. Everson, J. Meakin, I. L. 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.

[133] X. Zheng, Y. Wang, G. Wang, and J. Liu. Fast and robust segmentation of white blood cell images by self-supervised learning. *Micron*, 107:55–71, 2018.

[134] X. Zhuang, L. Li, C. Payer, D. Štern, M. Urschler, M. P. Heinrich, J. Oster, C. Wang, Ö. Smedby, C. Bian, et al. Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge. *Medical image analysis*, 58:101537, 2019.

[135] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao, and Y. J. Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our claims are justified through rigorous experiments evaluated on held-out datasets both quantitatively and qualitatively.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.

- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In addition to the core hypothesis of this work presented in Section 3, we have a dedicated limitations section at the end of the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: There are no theoretical results in this paper.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 3 has a dedicated infrastructure, training, and inference section that specifies parameters and the setup of the model development. The experimental setup is extensively discussed in Section 4. We provide additional information to reproduce the experiments and on model training in the supplemental material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The data we use comes from a collection of public datasets. They are all cited accordingly in the paper but we do not have authorization to release the data ourselves. Code will be made available upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not

including code, unless this is central to the contribution (e.g., for a new open-source benchmark).

- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All hyperparameters are specified in the main paper in Section 3 and 4 or in the supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All error bars are reported in the corresponding graphs and for each graph, we explain what they mean. The only graph were error bars are not directly visible is the heat map. We present the error bars in the supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).

- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: For the infrastructure experiment, we specify the compute resources details in Section 5. We also provide additional information in the supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We comply to the Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have a dedicated social impact section that discusses both.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our model is a segmentation model, not a pretrained LLM, image generator or scraped dataset. We do caution against the use of our model for out of distribution data.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Original owners of assets are credited for code, data and models.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

Justification: Commented and licensed code will be made public upon acceptance

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No human subjects or crowd-sourcing was involved.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No human subjects were involved.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLMs were only used for formatting, editing, and standard code cleaning and development.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Table of Content

We divide this supplemental material into five major parts:

Frequently asked questions (B). We address some questions that have been asked by colleagues about this work. We think this might be relevant to reviewers if this remains unclear in the paper.

Additional analysis on *Held-out* Datasets (C). We report results for the *Set Surface Distance* and *Set IoU* metrics. We analyze further architectural choices and their impact on performance. This includes the size of the set and the choice of M and K. We provide per-dataset performance when results are shown aggregated in the main paper.

Additional analysis on *Development* data (D). We present performance for the *Development* datasets, which were used to evaluate the generalization capabilities of the model.

Training infrastructure (E). We detail our training infrastructure and requirements.

Data (F). We provide additional information on Megamedical and Anatomix as well as the augmentations applied.

B Frequently Asked Questions

Instead of Pancakes, could a user fine-tune a model on a set of curated images? Pancakes tackles the prevalent scenario where there is no preexisting annotated data. It relieves the user of the requirement imposed by existing methods: annotating sufficient images to use for few-shot or in-context methods.

What is the difference between Pancakes and stochastic segmentation methods [10, 59]? Pancakes is focused on segmenting images in **new** tasks, not seen at training, while existing methods are specifically designed to model stochasticity in a given pre-specified task, essentially tackling a different problem. While Pancakes can also capture diversity for a fixed *single* protocol like existing models do, we model the substantially more challenging problem of presenting *different* plausible (multi-label) protocols for the new task.

What does it mean for segmentations to be consistent across a set? If a label ID appears across a set of anatomically similar images, this ID semantically corresponds to the same ROI in each image. For example, if label 1 represents the hippocampus in the first image, it should also represent the hippocampus in the second image.

Why is consistency important? Consistency is critical when the user wants to analyze the same set of structures for multiple images as is prevalent in population and longitudinal studies.

Is it possible to achieve consistency through post-processing if a method is inconsistent? For previous methods, matching segmentation maps across scans is not trivial for several reasons. Among them: (1) The segmentation maps in different scans are often so different that there are no clear labels to match. The same segmentation model could segment images on different bases and different images can contain different structures (2) Even when the same anatomical structure is segmented in two different images, they usually don't share the same label ID or don't have the same size as shown in Figure 2. This makes automated matching as post-processing step ill-posed.

How are the sets treated at training? We train with input tensors of dimension $B \times S \times C \times H \times W$, for batch B, set S, channel C, image height H and width W. We flatten these tensors to $(B \times S) \times C \times H \times W$ so that we can use architectures based on 2D-convolutions and 2D-UNets.

What is the intuition behind the min Dice loss? Using a regular expectation over the candidate label maps would lead the model to regress to the mean. This might be particularly harmful when there is high uncertainty around the set of plausible labels to output. Instead, using the best values leads to more diverse predictions. We refer to [96] for a more extensive comment on this type of loss.

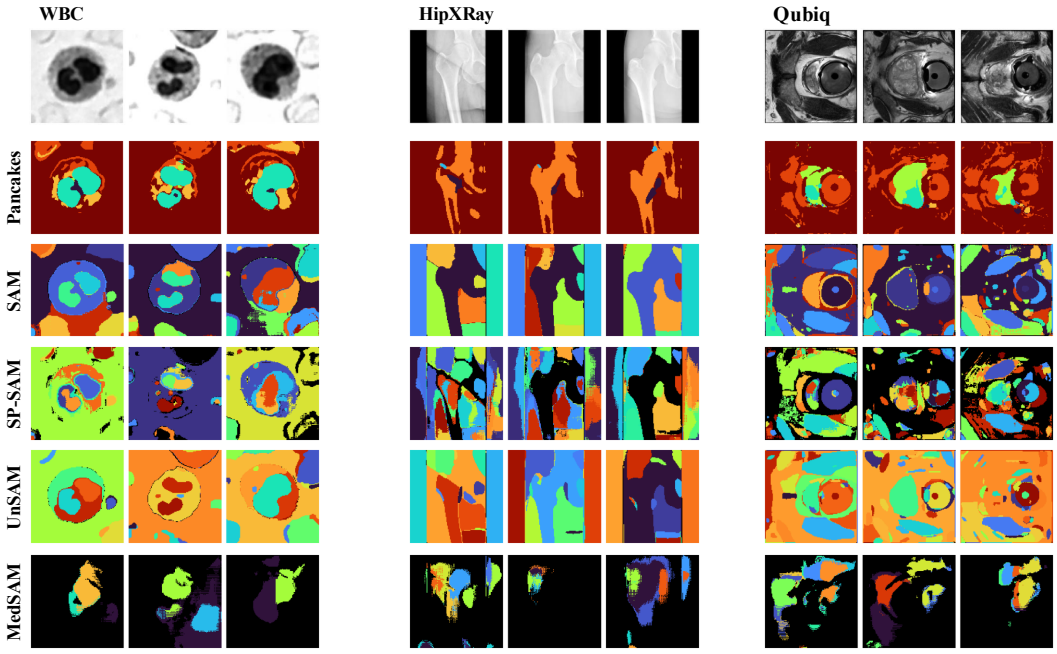


Figure 11: **Pancakes and baselines set consistency.** Evaluated on 3 held-out datasets (left to right: WBC, HipXRAY, QUBIQ Prostate), Pancakes is the method that is the most consistent across predictions compared to the baselines.

C Additional Analysis on *Held-out* Datasets

We present more in-depth results for the *held-out* datasets that remained unseen until the final phase of evaluation.

C.1 Qualitative comparison with baselines

We visually compare predictions for three held out datasets: WBC, HipXRay and QubiqProstate. Figure 11 shows that even though some baselines produce more accurate individual labels, Pancakes provides the most consistent labels across images of the same set.

C.2 Interpolation

This additional visualization explores the space of protocols M for multiple numbers of labels: K. For $M = 16$, we sample the 16 protocols available for different K values: $[5, 10, 15, 20, 25]$. We show in Figure 12 how this impacts the produced label maps for two samples in the WBC dataset. We find that segmentation maps that come from protocols *close* to one another in the embedding space are similar. Interestingly, the label space interpolated is relatively smooth, as indicated by the color and shape variations.

C.3 Additional Metrics

In Figure 13, we report results evaluated using two additional metrics: *Set IoU* and *Set surface distance*. They are computed in the same way as *Set Dice* but replacing Dice score with IoU and Surface Distance respectively. For the surface distance, we keep the percentile 95. When the set size is one ($S = 1$), the set metric reduces to the standard metric, ignoring consistency.

C.4 Per set size performance

We report additional per-set-size results that were summarized in the main paper.

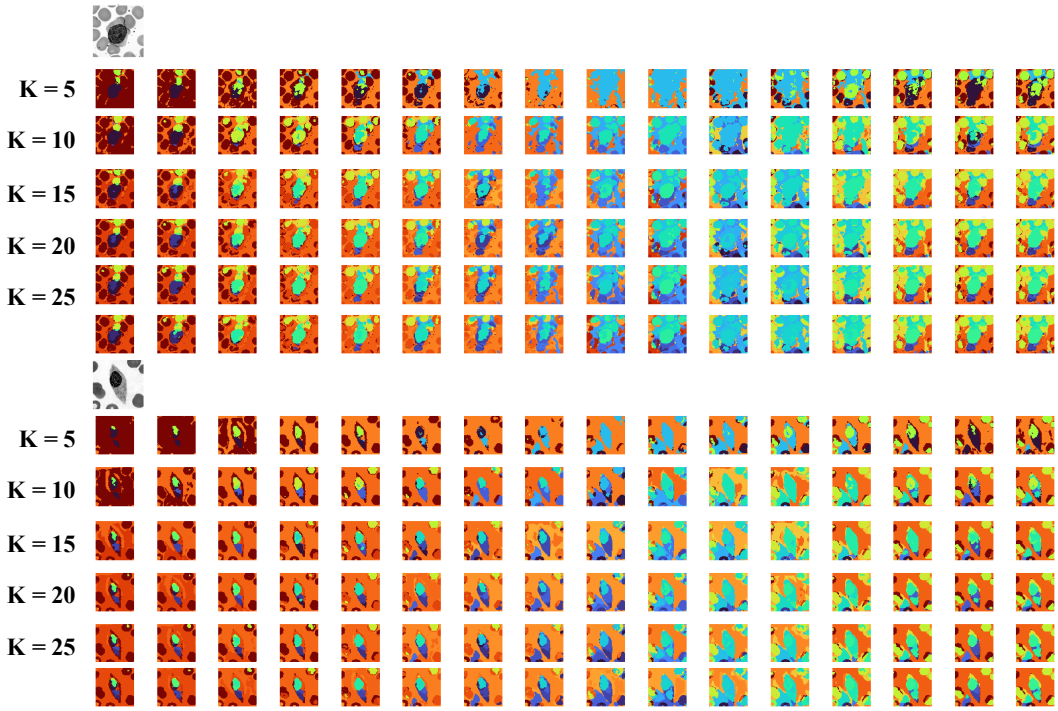


Figure 12: **Interpolation Analysis.** By covering the space of protocol (rows) for several fixed K values, we observe that label maps resulting from protocols *close* to another in the embedding space are similar.

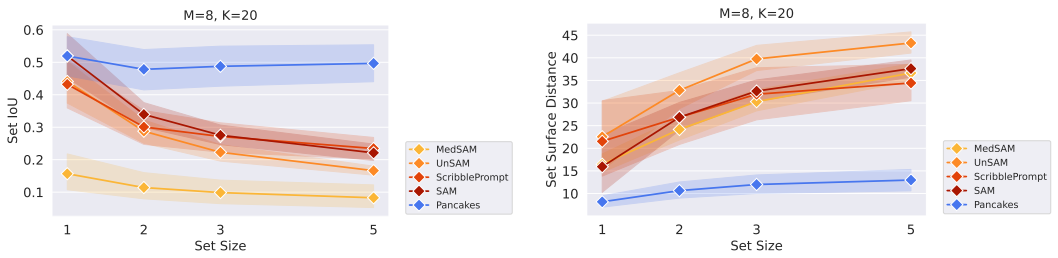
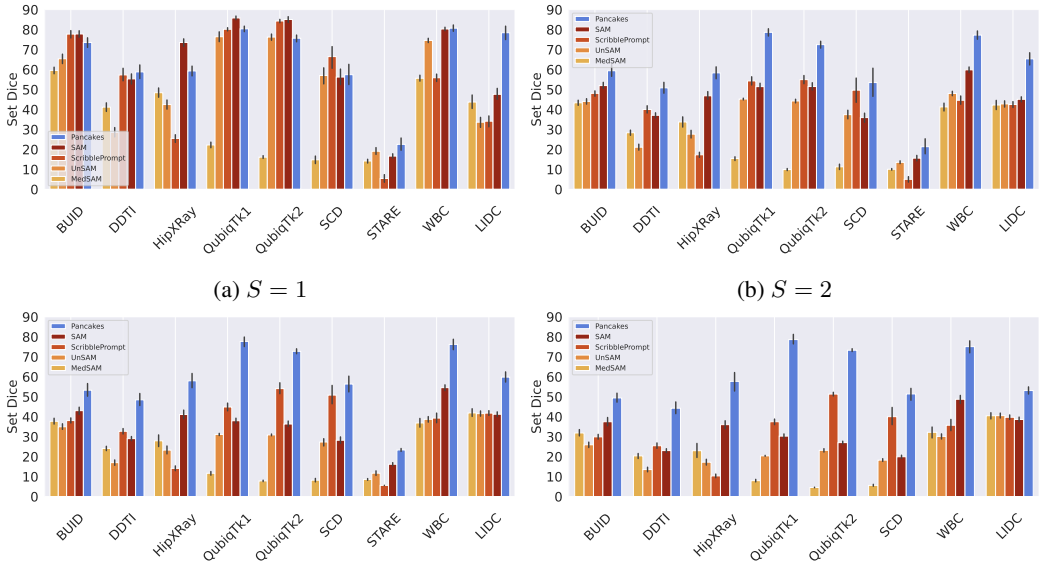


Figure 13: **Evaluation on additional metrics: IoU and surface distance.** For individual predictions ($S = 1$), Pancakes is comparable to SAM and outperforms baselines. As the set size S increases, Pancakes is the only model whose performance is not severely degraded, as it can produce consistent segmentations.



(a) $S = 1$ (b) $S = 2$

(c) $S = 3$ (d) $S = 5$

Figure 14: **Per-dataset *Set Dice* performance under various set sizes.** We evaluate Pancakes and baselines with various set sizes: 1, 2, 3 and 5. For Pancakes, we evaluate with $M = 8$ and $K = 20$.

Per-dataset *Set Dice* performance. This experiment evaluates how segmentation accuracy and label consistency vary jointly with set size. Figure 14 shows that Pancakes's improvement over the baselines increases with the set size. Figure 14a focuses on accuracy only as the set size is one ($S = 1$). In this case, Pancakes is comparable to the baselines (Figure 14a). We observe a large variability across different datasets.

Influence of M and K. Figure 15 shows the influence of the number of protocols and labels on the Pancakes predictions as the set size varies. As the number of protocols increases, performance increases. Best performing K seems to be in the middle of the range for K=20. Results are consistent across set sizes.

D Additional Analysis on *Development* Datasets

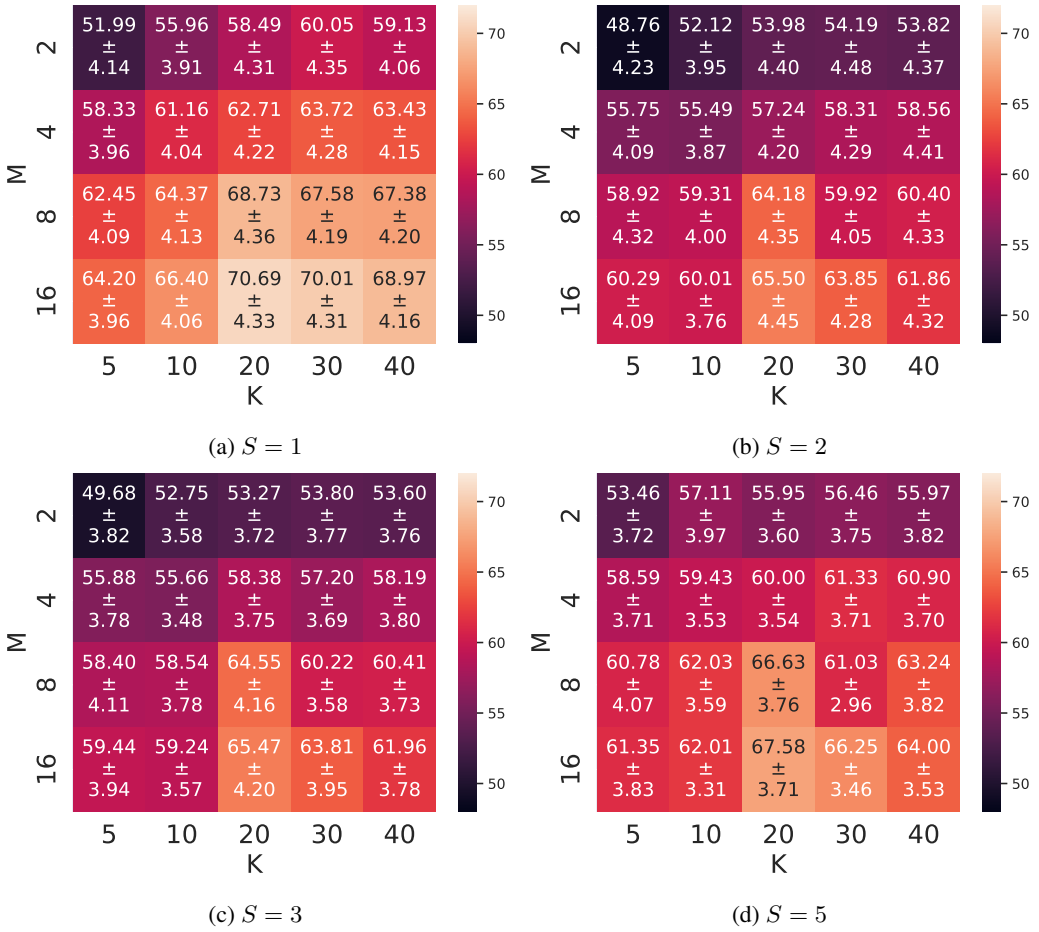
We include performance on the *Development* datasets that were used to evaluate generalization capabilities during model development: ACDC [11], PanDental [2] and SpineWeb [132].

We evaluate on the test split of all datasets. When reporting performance with a certain set size, we exclude tasks with subjects fewer than this set size. (For example, the test split of SpineWeb has only 2 subjects, so we do not include it in set size 3 and 5 experiments.) We report Pancakes performance with $M = 8$ and $K = 20$. Figure 16 shows Pancakes demonstrates both accurate and consistent predictions compared to the baselines.

Per-Dataset *Set Dice* Performance for Different Set Sizes. Figure 17 shows per-dataset performance, separated by set size.

Influence of M and K. Figure 18 shows effect of number of protocols and labels per protocol on prediction accuracy, as set size varies. The trend aligns with results using the held-out datasets (C.4).

Learning a complete protocol. The primary objective of Pancakes is to propose a diverse set of consistent protocols for an image group. We do not enforce constraints across protocols produced, but at times it would be good to do so. For example, symmetric labels (left and right ventricles or posterior and anterior hippocampus) in the same protocol may be desirable. One way to incorporate this into the Pancakes framework is to apply the loss function to two randomly chosen segmentation



(a) $S = 1$

(b) $S = 2$

(c) $S = 3$

(d) $S = 5$

Figure 15: **Influence of varying M and K for various set sizes on the *Held-out* datasets.** We evaluate Pancakes with various set sizes, number of protocols (M) and number of labels in each protocol (K).

targets in a multi-label dataset, rather than one. As a preliminary experiment, we ran an experiment on the OASIS brain dataset, and found it effective at learning a protocol as shown in Figure 19. We are planning to explore this more in future work.

E Training Infrastructure

Our model was trained using 45G of memory on a single node of an NVIDIA DGX A100 machine using two cores. We use a batch size of 1 and the AdamW optimizer with a learning rate of 0.0001. We use PReLU activations and convolution layers with 32 features, kernel size 3 and stride 1.

F Data

F.1 Megamedical

We train our main experiment on Megamedical [16, 96, 123]. The images are 2D and resized to 128×128. Complete tables of the datasets, split by data subgroup (*Training*, *Development*, *Held-out*) are shown in Tables 4, 5, and 6. Megamedical covers a wide range of modalities (MRI, ultra-sound, XRay) and anatomies (organs, bones, substructures and fine structures like vessels).

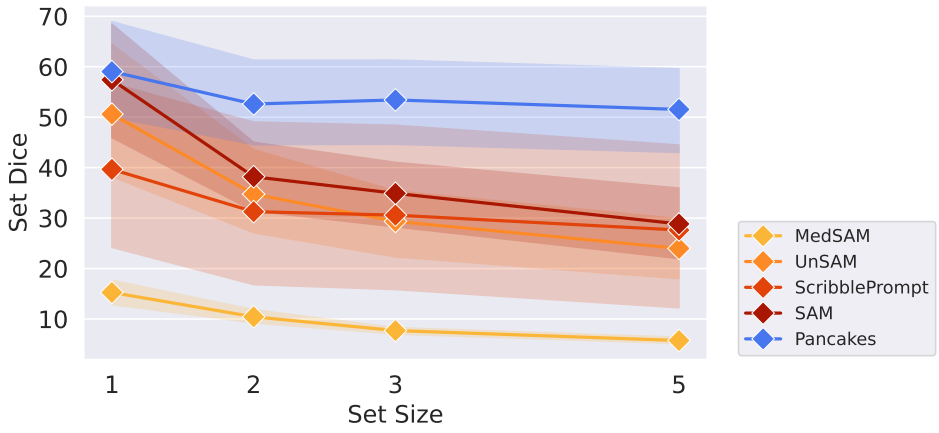
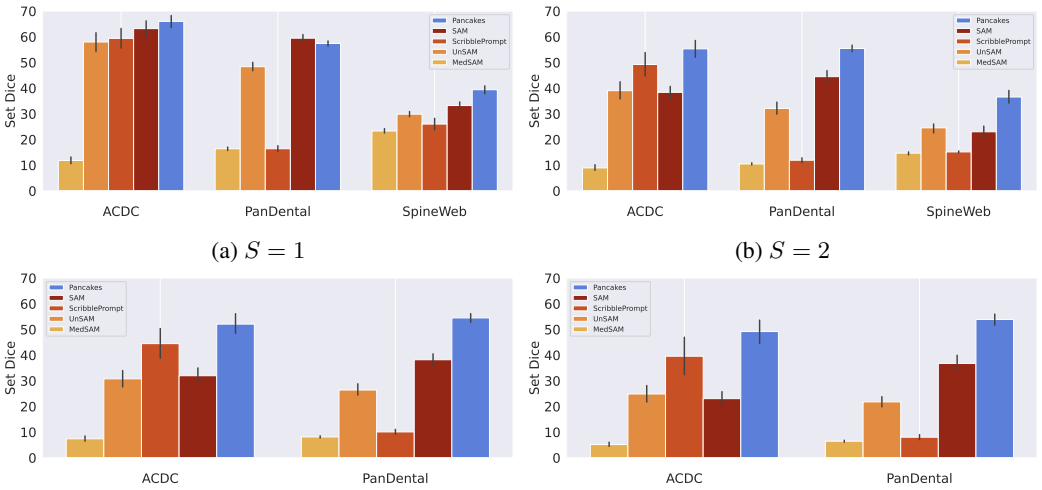


Figure 16: **Pancakes is both consistent and accurate compared to baselines.** We evaluate Pancakes and the baselines on *Development* datasets: ACDC, PanDental and SpineWeb. When evaluated solely on accuracy ($S = 1$), Pancakes is comparable to SAM and outperforms the other baselines. As S increases, Pancakes is the only model whose performance is not degraded.



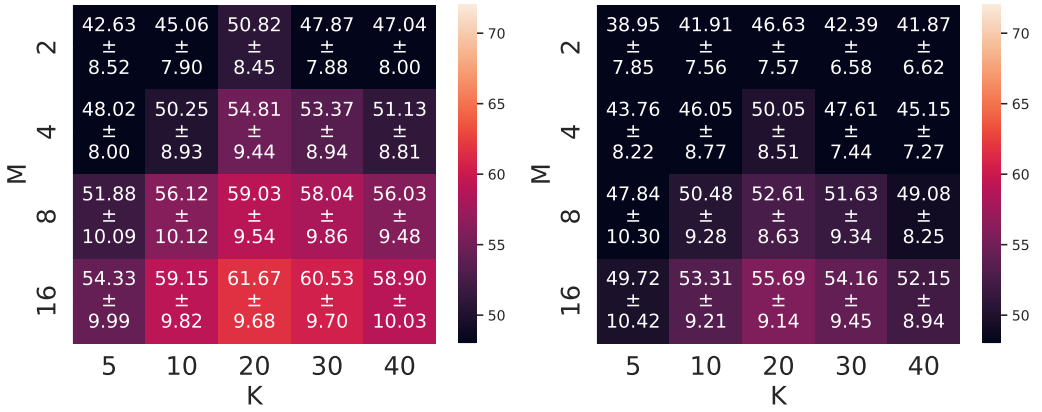
(a) $S = 1$ (b) $S = 2$

(c) $S = 3$ (d) $S = 5$

Figure 17: **Per-dataset *Set Dice* performance under various set sizes for the *Development* datasets.** We evaluate Pancakes and baselines on development datasets (ACDC, PanDental and SpineWeb) with various set sizes: 1, 2, 3 and 5. If a dataset has fewer subjects than the set size, we exclude it. For Pancakes, we use $M = 8$ and $K = 20$.

F.2 Anatomix

Building on [13, 16, 43] and specifically on Anatomix [27], we use synthetic data to complement Megamedical and limit overfitting. To generate Anatomix label maps, we sample a random set of 3D labels from the TotalSegmentator dataset [118]. We sample between 20 and 40 labels and generate a $128 \times 128 \times 128$ label map. Once this 3D label map is generated, we randomly sample an axis and then a slice between slice ID 25 and 100. With probability 50%, we will *split* labels. In that case, labels are reassigned so that a given label can only be composed of contiguous pixels. We also assign to background (label ID 0) any label whose size is smaller than 20 pixels. This gives us the label map template for a given set, from which we are going to generate the images and label maps for each element in the set. We generate the images by randomly assigning an intensity value to each label. We then apply a series of augmentations that are independent for each element in the set. These augmentations include: Gaussian blur, Gaussian noise, Perlin noise, elastic and affine transforms,



(a) $S = 1$ (b) $S = 2$

(c) $S = 3$ (d) $S = 5$

Figure 18: **Influence of varying M and K in various set sizes for the *Development* datasets.** On development datasets, we evaluate Pancakes with various set sizes, number of protocols (M) and number of labels in each protocol (K).

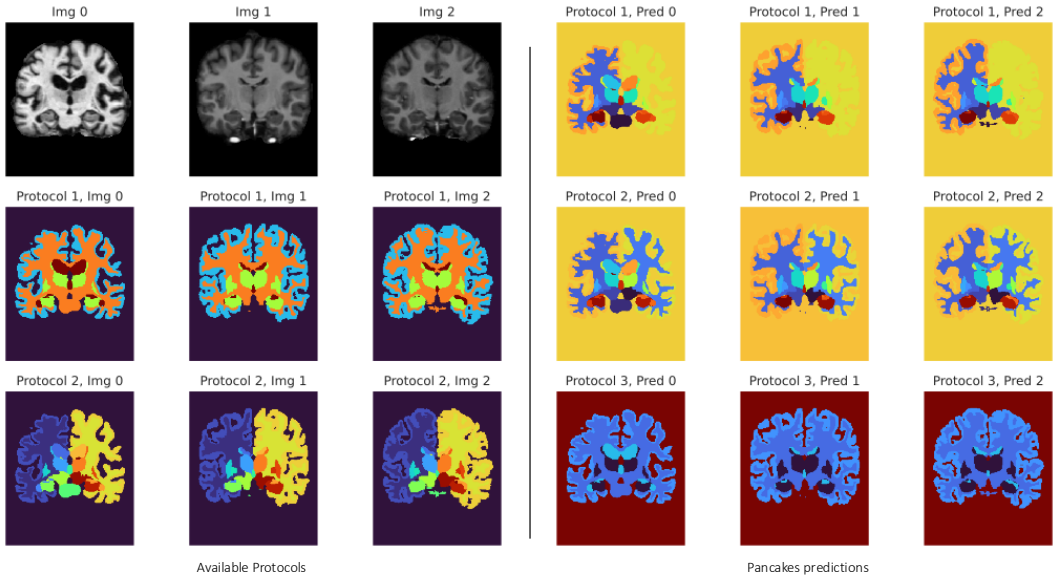


Figure 19: **Learning a full protocol.** Training on OASIS, we learn accurately a complete protocol with our loss modification.

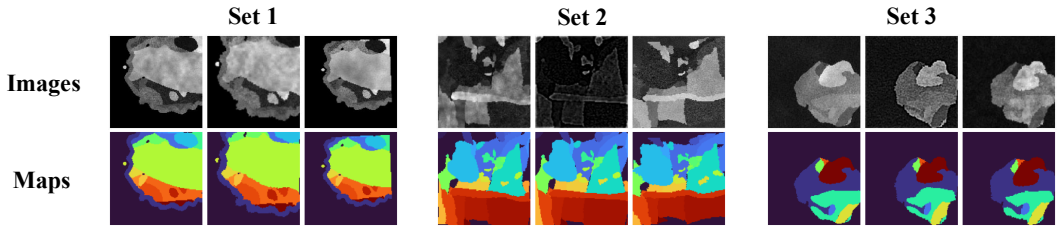


Figure 20: **Anatomix Examples.** Label maps and images within a set are very similar to one another, sharing similar relative locations for each structure. The main difference lies in the augmentations applied, to the images individually (Gaussian noise for example), but also to both the images and label maps (elastic deformation for example).

and contrast variations. To sample a binary ground truth label at training, we sample a label that is present in each image of the set. Example sets generated are shown in Figure 20.

F.3 Data Augmentation

At training, we apply a series of augmentations to our set to improve generalization capabilities. We distinguish between two types of augmentations, *within-set* augmentation — applied independently to each element in the set — and *across-set* augmentation — applied consistently to each element in the set. For each augmentation sampled, the corresponding parameters are sampled uniformly from a pre-defined range. Table 3 shows the list of all augmentations applied, the probability of each augmentation per iteration and the parameter ranges we sample from.

Table 3: **Augmentations used to train** We apply augmentations either independently within a set (Top) or to all the elements of a set (Bottom). We sample each parameter from the uniform distribution ($\mathcal{U}$) within the ranges defined in the *Parameters* column.

(a) Within-Set Augmentation

Augmentations	p	Parameters
Random Affine	0.25	degrees $\sim \mathcal{U}(-25, 25)$ translate$\sim \mathcal{U}(0, 0.1)$ scale$\sim \mathcal{U}(0.9, 1.1)$
Brightness Contrast	0.5	brightness$\sim \mathcal{U}(-0.1, 0.1)$, contrast $\sim \mathcal{U}(0.5, 1.5)$
Elastic Transform	0.8	$\alpha \sim \mathcal{U}(1, 2.5)$ $\sigma \sim \mathcal{U}(7, 9)$
Sharpness	0.25	sharpness$= 3$
Flip Intensities	0.5	None
Gaussian Blur	0.25	$\sigma \sim \mathcal{U}(0.1, 1)$ k=5
Gaussian Noise	0.25	$\mu \sim \mathcal{U}(0, 0.05)$ $\sigma \sim \mathcal{U}(0, 0.05)$

(b) Across-Set Augmentation

Augmentations	p	Parameters
Random Affine	0.5	degrees $\sim \mathcal{U}(0, 360)$ translate$\sim \mathcal{U}(0, 0.2)$ scale$\sim \mathcal{U}(0.8, 1.1)$
Brightness Contrast	0.5	brightness$\sim \mathcal{U}(-0.1, 0.1)$, contrast$\sim \mathcal{U}(0.8, 1.2)$
Gaussian Blur	0.5	$\sigma \sim \mathcal{U}(0.1, 1.1)$ $k = 5$
Gaussian Noise	0.5	$\mu \sim \mathcal{U}(0, 0.05)$ $\sigma \sim \mathcal{U}(0, 0.05)$
Elastic Transform	0.5	$\alpha \sim \mathcal{U}(1, 2)$ $\sigma \sim \mathcal{U}(6, 8)$
Sharpness	0.5	sharpness$= 5$
Horizontal Flip	0.5	None
Vertical Flip	0.5	None

Table 4: **Collection of datasets in Megamedical used for training**. The entry # of scans is the number of unique (subject, modality) pairs for each dataset.

Dataset Name	Description	# of Scans	Image Modalities
AMOS [49]	Abdominal organ segmentation	240	CT, MRI
BBBC003 [72]	Mouse embryos	15	Microscopy
BBBC038 [17]	Nuclei images	670	Microscopy
BrainDev. [33, 34, 63, 104]	Adult and Neonatal Brain Atlases	53	multi-modal MRI
BRATS [6, 7, 84]	Brain tumors	6,096	multi-modal MRI
BTCV [65]	Abdominal Organs	30	CT
BUS [128]	Breast tumor	163	Ultrasound
CAMUS [66]	Four-chamber and Apical two-chamber heart	500	Ultrasound
CDemris [51]	Human Left Atrial Wall	60	CMR
CHAOS [53, 55]	Abdominal organs (liver, kidneys, spleen)	40	CT, T2-weighted MRI
CheXplanation [101]	Chest X-Ray observations	170	X-Ray
CT-ORG[98]	Abdominal organ segmentation (overlap with LiTS)	140	CT
DRIVE [110]	Blood vessels in retinal images	20	Optical camera
EOphtha [25]	Eye Microaneurysms and Diabetic Retinopathy	102	Optical camera
FeTA [89]	Fetal brain structures	80	Fetal MRI
FetoPlac [9]	Placenta vessel	6	Fetoscopic optical camera
HMC-QU [26, 57]	4-chamber (A4C) and apical 2-chamber (A2C) left wall	292	Ultrasound
I2CVB [67]	Prostate (peripheral zone, central gland)	19	T2-weighted MRI
IDRID [93]	Diabetic Retinopathy	54	Optical camera
ISLES [42]	Ischemic stroke lesion	180	multi-modal MRI
KiTS [41]	Kidney and kidney tumor	210	CT
LGGFlair [15, 82]	TCIA lower-grade glioma brain tumor	110	MRI
LiTS [12]	Liver Tumor	131	CT
LUNA [105]	Lungs	888	CT
MCIC [32]	Multi-site Brain regions of Schizophrenic patients	390	T1-weighted MRI
MSD [106]	Collection of 10 Medical Segmentation Datasets	3,225	CT, multi-modal MRI
NCI-ISBI [14]	Prostate	30	T2-weighted MRI
OASIS [45, 80]	Brain anatomy	414	T1-weighted MRI
OCTA500 [69]	Retinal vascular	500	OCT/OCTA
PAXRay [103]	Thoracic organs	880	X-Ray
PROMISE12 [71]	Prostate	37	T2-weighted MRI
PPMI [81]	Brain regions of Parkinson patients	1,130	T1-weighted MRI
QUBIQ [83]	Brain, kidney, pancreas	209	MRI T1, Multimodal MRI, CT
ROSE [78]	Retinal vessel	117	OCT/OCTA
SegTHOR [64]	Thoracic organs (heart, trachea, esophagus)	40	CT
ToothSeg [47]	Individual teeth	598	X-Ray
WMH [62]	White matter hyper-intensities	60	multi-modal MRI
WORD [75]	Organ segmentation	120	CT

Table 5: **Development datasets**. Datasets used to evaluate the generalization capabilities of our model and model development.

Dataset Name	Description	# of Scans	Image Modalities
ACDC [11]	Left and right ventricular endocardium	99	cine-MRI
PanDental [2]	Mandible and Teeth	215	X-Ray
SpineWeb [132]	Vertebrae	15	T2-weighted MRI

Table 6: **Held-out datasets**. Datasets that remained unseen until the final evaluation phase.

Dataset Name	Description	# of Scans	Image Modalities
BUID [3]	Breast tumors	647	Ultrasound
DDTI [91]	Thyroid	472	Ultrasound
HipXRay [37]	Ilium and femur	140	X-Ray
QUBIQ [83]	Prostate	209	MRI T1, Multimodal MRI, CT
SCD [94]	Sunnybrook Cardiac Multi-Dataset Collection	100	cine-MRI
LIDC-IDRI [5]	Lung Nodules	1,018	CT
STARE [46]	Blood vessels in retinal images	20	Optical camera
WBC [133]	White blood cell and nucleus	400	Microscopy