UniverSeg: Universal Medical Image Segmentation

Victor Ion Butoi*
MIT CSAIL
vbutoi@mit.edu

Jose Javier Gonzalez Ortiz*
MIT CSAIL
josejg@mit.edu

Tianyu Ma
Cornell University
tm478@cornell.edu

Mert R. Sabuncu
Cornell University
msabuncu@cornell.edu

John Guttag
MIT CSAIL
guttag@mit.edu

Adrian V. Dalca
MIT CSAIL & MGH, HMS
adalca@mit.edu

Abstract

While deep learning models have become the predominant method for medical image segmentation, they are typically not capable of generalizing to unseen segmentation tasks involving new anatomies, image modalities, or labels. Given a new segmentation task, researchers generally have to train or fine-tune models. This is time-consuming and poses a substantial barrier for clinical researchers, who often lack the resources and expertise to train neural networks.

We present UniverSeg, a method for solving unseen medical segmentation tasks without additional training. Given a query image and an example set of image-label pairs that define a new segmentation task, UniverSeg employs a new CrossBlock mechanism to produce accurate segmentation maps without additional training. To achieve generalization to new tasks, we have gathered and standardized a collection of 53 open-access medical segmentation datasets with over 22,000 scans, which we refer to as MegaMedical. We used this collection to train UniverSeg on a diverse set of anatomies and imaging modalities. We demonstrate that UniverSeg substantially outperforms several related methods on unseen tasks, and thoroughly analyze and draw insights about important aspects of the proposed system. The UniverSeg source code and model weights are freely available at https://universeg.csail.mit.edu

*Denotes equal contribution

1. Introduction

Image segmentation is a widely studied problem in computer vision and a central challenge in medical image analysis. Medical segmentation tasks can involve diverse imaging modalities, such as magnetic resonance imaging (MRI), X-ray, computerized tomography (CT), and microscopy; different biomedical domains, such as the abdomen, chest, brain, retina, or individual cells; and different labels within a region, such as heart valves or chambers (Figure 1). This diversity has inspired a wide array of segmentation tools, each usually tackling one task or a small set of closely related tasks [17, 23, 41, 42, 87, 94]. In recent years, deep-learning models have become the predominant strategy for medical image segmentation [45, 74, 87].

A key problem in image segmentation is *domain shift*,

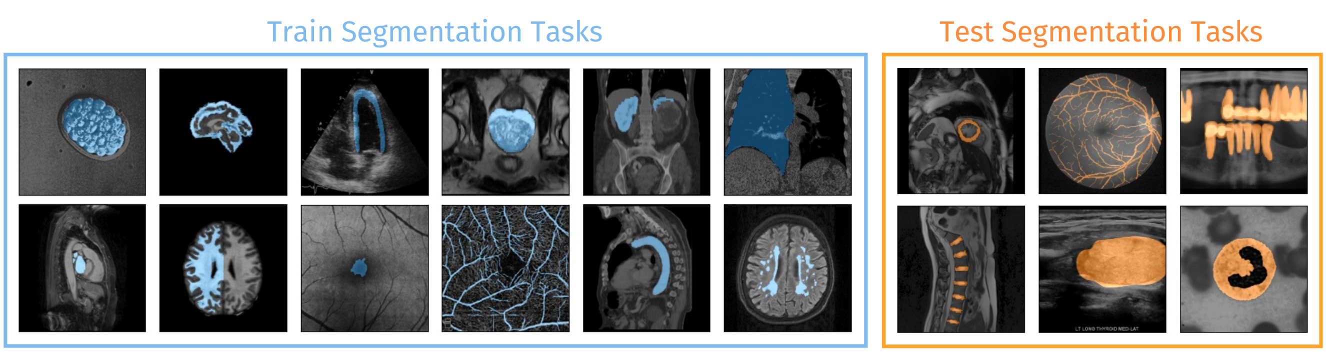

Figure 1: Medical segmentation involves many imaging types, biomedical domains, and target labels. We employ a large diverse set of training tasks **(blue)** to build a model that can segment unseen tasks **(orange)** without additional training.

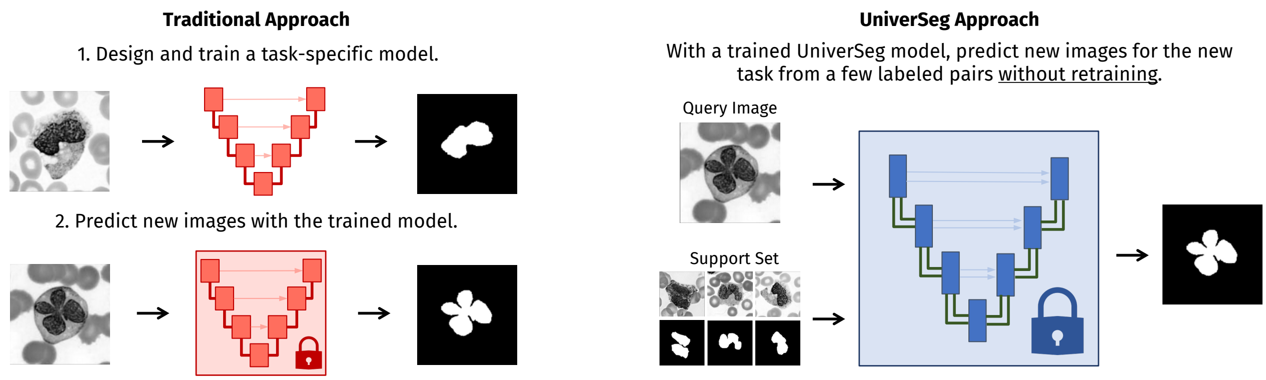

Figure 2: **Workflow for inference on a new task, from an unseen dataset.** Given a new task, traditional models **(left)** are trained before making predictions. UniverSeg **(right)** employs a *single* trained model which can make predictions for images (queries) from the new task with a few labeled examples as input (support set), without additional fine-tuning.

where models often perform poorly given out-of-distribution examples. This is especially problematic in the medical domain where clinical researchers or other scientists are constantly defining new segmentation tasks driven by evolving populations, and scientific and clinical goals. To solve these problems they need to either train models from scratch or fine-tune existing models. Unfortunately, training neural networks requires machine learning expertise, computational resources, and human labor. This is infeasible for most clinical researchers or other scientists, who do not possess the expertise or resources to train models. In practice, this substantially slows scientific development. We, therefore, focus on avoiding the need to do *any* training given a new segmentation task.

Fine-tuning models trained on the natural image domain can be unhelpful in the medical domain [86], likely due to the differences in data sizes, features, and task specifications between domains, and importantly still requires substantial retraining. Some few-shot semantic segmentation approaches attempt to predict novel classes without fine-tuning in limited data regimes, but mostly focus on classification tasks, or segmentation of new classes within the same input domain, and do not generalize across anatomies or imaging modalities.

In this paper, we present UniverSeg – an approach to learning a *single* general medical-image segmentation model that performs well on a variety of tasks without any retraining, including tasks that are substantially different from those seen at training time. UniverSeg learns how to exploit an input set of labeled examples that specify the segmentation task, to segment a new biomedical image in one forward pass. We make the following contributions.

- We propose UniverSeg – a framework that enables solving new segmentation tasks without retraining, using a novel flexible CrossBlock mechanism that transfers information from the example set to the new image.
- We demonstrate that UniverSeg substantially outperforms several models across diverse held-out segmentation tasks involving unseen anatomies and even approaches the performance of fully-supervised networks trained specifically for those tasks.
- In extensive analysis, we show that the generalization capabilities of UniverSeg are linked to task diversity during training and image diversity during inference.

2. Related Works

Medical Image Segmentation. Medical image segmentation has been widely studied, with state-of-the-art methods training convolutional neural networks in a supervised fashion, predicting a label map for a given input image [23, 41, 42, 46, 87]. For a new segmentation problem, models are typically trained from scratch, requiring substantial design and tuning.

Recent strategies, such as the nnUNet [42], automate some design decisions such as data processing or model architecture but still incur substantial overhead from training. In contrast to these methods, UniverSeg generalizes to new medical segmentation tasks without training or fine-tuning.

Multi-task Learning. Multi-Task Learning (MTL) frameworks learn several tasks simultaneously [16, 24, 90]. For medical imaging, this can involve multiple modalities [75], population centers [64], or anatomies [76]. However, the tasks are always pre-determined by design: once trained, each network can only solve tasks presented during training. UniverSeg overcomes this limitation, enabling tasks to be dynamically specified during inference.

Transfer Learning. Transfer learning strategies involve fine-tuning pre-trained models, often from a different domain [66, 101]. This is used in medical image segmentation starting

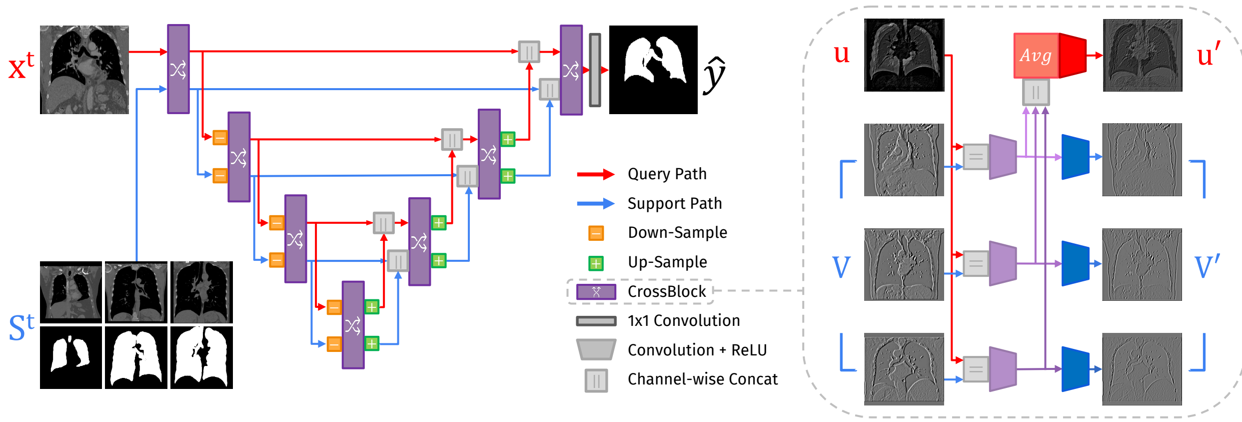

Figure 3: A UniverSeg network **(left)** takes as input a query image and a support set of image and label-maps (pairwise concatenated in the channel dimension) and employs multi-scale CrossBlock features. A CrossBlock **(right)** takes as input representations of the query u and support set $V = \{v_i\}$, and interacts u with each support entry v_i to produce u' and V'.

with models trained on natural images [4, 27, 44, 113, 116], where the amount of data far exceeds the amount in the target biomedical domain. However, this technique still involves substantial training for each new task, which UniverSeg avoids. Additionally, the differences between medical and natural images often make transfer learning from large pre-trained models unhelpful [86].

Optimization-based Meta-Learning. Optimization-based meta-learning techniques often learn representations that minimize downstream fine-tuning steps by using a few examples per task, sometimes referred to as few-shot learning [25, 78, 98, 104]. Meta-learning via fine-tuning has been studied in medical image segmentation to handle multiple image modalities [112], anatomies [110], and generalization to different targets [51, 52, 97]. While these strategies reduce the amount of data and training required for downstream tasks [33], fine-tuning these models nevertheless requires machine learning expertise and computational resources, which are often not available to medical researchers.

In-Context Learning. In-Context Learning (ICL) methods adapt to new tasks without additional training by incorporating the task description as an input to the model [14]. This strategy has been successfully demonstrated in both large language models [79, 108] and multi-modal foundation models which take interleaved text and images as inputs, maintaining natural language as the primary prompting mechanism [3]. Differing from recent image-based ICL image models which employ transformers [8, 106, 107], we develop a purely convolutional in-context learning method for medical image segmentation tasks, in which tasks are encoded as sets of image-label pairs.

Few-shot Semantic Segmentation. Few-shot (FS) methods adapt to new tasks from few training examples, often by fine-tuning pretrained networks [25, 78, 104, 98]. Some few-shot semantic segmentation models generate predictions for new images (queries) containing unseen classes from just a few labeled examples (support) without additional retraining. One strategy prevalent in both natural image [77, 91, 109] and medical image [22, 62, 81, 95] FS segmentation methods is to employ large pre-trained models to extract deep features from the query and support images. These methods often involve learning meaningful prototypical representations for each label [105]. Self-supervised learning can help make up for the lack of training data and tasks [32, 80]. In contrast to UniverSeg, these methods focus on limited data regimes, tackle specific tasks, and generalize to new classes in a particular subdomain, like abdominal CT or MRI scans [32, 80, 88, 102]. In our work, we focus on avoiding *any* fine-tuning, even when given many examples for a new task, to avoid requiring the clinical or scientific user to have machine learning expertise and computing resources.

3. UniverSeg Method

Let t be a segmentation task comprised of a set of image-label pairs $\{(x_i^t, y_i^t)\}_{i=1}^N$. Common segmentation strategies learn parametric functions $\hat{y} = f_\theta^t(x)$, where f_θ^t is most often modeled using a convolutional neural network that estimates a label map $\hat{y}$ given an input image x. By construction, f_θ^t only learns to predict segmentations for task t.

In contrast, we learn a universal function $\hat{y} = f_\theta(x^t, S^t)$ that predicts a label map for input x^t of task t, according to the task-specifying support $S^t = \{(x_j^t, y_j^t)\}_{j=1}^n$ comprised of example image-label pairs available for t.

3.1. Model

We implement f_θ using a fully convolutional neural network illustrated in Figure 3. We first introduce the proposed building blocks: the *cross-convolution* layer and the CrossBlock module. We then specify how we combine these blocks into a complete segmentation network.

CrossBlock. To transfer information between the support set and query image, we introduce a *cross-convolution* layer that interacts a query feature map u with a set of support feature maps $V = \{v_i\}_{i=1}^n$:

$$\begin{aligned} \text{CrossConv}(u, V; \theta_z) &= \{z_i\}_{i=1}^n, \\ \text{for } z_i &= \text{Conv}(u||v_i; \theta_z), \end{aligned} \tag{1}$$

where $||$ is the concatenation operation along the feature dimension and $\text{Conv}(x; \theta_z)$ is a convolutional layer with learnable parameters θ_z. Due to the weight reuse of θ_z, cross-convolution operations are permutation invariant with respect to V. From this layer, we design a higher-level building block that produces updated versions of query representation u and support V at each step in the network:

$$\begin{aligned} &\text{CrossBlock}(u, V; \theta_z, \theta_u, \theta_v) = (u', V'), \text{where:} \qquad (2)\\ &\quad z_i = \phi(\text{CrossConv}(u, v_i; \theta_z)) \quad \text{for } i = 1, 2, \ldots, n \\ &\quad u' = \phi(\text{Conv}(1/n \textstyle\sum_{i=1}^n z_i; \theta_u)) \\ &\quad v_i' = \phi(\text{Conv}(z_i; \theta_v)) \quad \text{for } i = 1, 2, \ldots, n, \end{aligned}$$

where $\phi(x)$ is a non-linear activation function. This strategy enables the representations of each support set entry and query to interact with the others through their average representation, and facilitates variably sized support sets.

Network. To integrate information across spatial scales, we compose the CrossBlock modules in an encoder-decoder structure with residual connections, similarly to the popular UNet architecture (Figure 3). The network takes as input the query image x^t and support set $S^t = \{(x_i^t, y_i^t)\}_{i=1}^n$ of image and label-map pairs, each concatenated channel-wise, and outputs the segmentation prediction map $\hat{y}^t$.

Each level in the encoder path consists of a CrossBlock followed by a spatial down-sampling operation of both query and support set representations. Each level in the expansive path consists of up-sampling both representations, which double their spatial resolutions, concatenating them with the equivalently-sized representation in the encoding path, followed by a CrossBlock. We perform a single 1x1 convolution to map the final query representation to a prediction.

3.2. Training

Algorithm 1 describes UniverSeg training using a large and varied set of training tasks $\mathcal{T}$ and the loss

$$\mathcal{L}(\theta; \mathcal{T}) = \mathbb{E}_{t \in \mathcal{T}} \mathbb{E}_{(x^t, y^t), S^t} \Big[\mathcal{L}_{\text{seg}}(f_\theta(x^t, S^t), y^t) \Big], \tag{3}$$

where $x^t \notin S^t$, and $\mathcal{L}_{\text{seg}}(\hat{y}, y^t)$ is a standard segmentation loss like cross-entropy or soft Dice [74], capturing the agreement between the predicted $\hat{y}$ and ground truth y_t.

Data Augmentation. We employ data augmentation to grow the diversity of training tasks and increase the number of effective training examples belonging to any particular task.

In-Task Augmentation – $\text{Aug}_t(x, y)$. To reduce overfitting to individual subjects, we perform standard data augmentation operations, like affine transformations, elastic deformation, or adding image noise to the query image and *each entry* of the support set independently.

Task Augmentation – $\text{Aug}_T(x, y, S)$. Similar to standard data augmentation that reduces overfitting to training examples, augmenting the training *tasks* is useful for generalizing to *new tasks*, especially those far from the training task distribution. We introduce task augmentation – alterations that modify all query and support images, and/or all segmentation maps, with the same type of task-changing transformation. Example task augmentations include edge detection of the segmentation maps or a horizontal flip to all images and labels. We provide a list of all augmentations and the parameters we used in the supplemental Section C.

Algorithm 1 UniverSeg Training Loop using SGD with learning rate η over tasks $\mathcal{T}$, main architecture f_θ, in-task augmentations Aug_t and task augmentations Aug_T

```
for k = 1, ..., NumTrainSteps do
    t ~ T                                        ▷ Sample Task
    (x_i^t, y_i^t) ~ t                           ▷ Sample Query
    S^t ← {(x_j^t, y_j^t)}_{j≠i}^n               ▷ Sample Support
    x_i^t, y_i^t ← Aug_t(x_i^t, y_i^t)           ▷ Augment Query
    S^t ← {Aug_t(x_j^t, y_j^t)}_j^n              ▷ Augment Support
    x_i^t, y_i^t, S^t ← Aug_T(x_i^t, y_i^t, S^t) ▷ Task Aug
    ŷ_i ← f_θ(x_i^t, S^t)                        ▷ Predict label map
    ℓ ← L_seg(ŷ_i, y_i^t)                        ▷ Compute loss
    θ ← θ − η∇_θ ℓ                               ▷ Gradient step
end for
```

3.3. Inference

For a given query image x^t, UniverSeg predicts segmentation $\hat{y} = f_\theta(x^t, S^t)$ given a support set S^t , where the prediction quality depends on the choice of the support set S^t. To reduce this dependence, and to take advantage of more data when memory constraints limit the support set size at inference, we combine predictions from an ensemble of K independently sampled support sets $\{S_i^t\}_{i=1}^K$ as their the pixel-wise average to produce the prediction $\hat{y} = \frac{1}{K} \sum_{k=1}^K f_\theta(x, S_k^t)$.

4. MegaMedical Dataset

To train our universal model f_θ, we employ a set of segmentation tasks that is large and diverse, so that it is able to generalize to new tasks. We compiled MegaMedical – an extensive collection of open-access medical segmentation datasets with diverse anatomies, imaging modalities, and labels. It is constructed from 53 datasets encompassing 26 medical domains and 16 imaging modalities.

We standardize data across the wildly diverse formats of original datasets, processed images, and label maps. We also expand the training data using synthetic segmentation tasks to further increase the training task diversity. Because of individual dataset agreements, we are prohibited from re-releasing our processed version of the datasets. Instead, we will provide data processing code to construct MegaMedical from its source datasets.

Datasets. MegaMedical features a wide array of biomedical domains, such as eyes [40, 61, 69, 84, 99], lungs [89, 93, 96], spine vertebrae [114], white blood cells [115], abdominal [11, 13, 35, 43, 49, 57, 58, 60, 63, 67, 68, 85, 96], and brain [5, 28, 36, 55, 56, 70, 71, 72, 96], among others. Supplemental Table 3 provides a detailed list of MegaMedical datasets. Acquisition details, subject age ranges, and health conditions are different for each dataset. We provide data processing details in supplemental Section A.

Medical Image Task Creation. While datasets in MegaMedical feature a variety of imaging tasks and label protocols, in this work we focus on the general problem of 2D binary segmentation. For datasets featuring 3D data, for each subject, we extract the 2D mid-slice of the volume along all the major axes. When multiple modalities are present, we include each modality as a new task. For datasets containing multiple segmentation labels, we create as many binary segmentation tasks as available labels. All images are resized to 128×128 pixels and intensities are normalized to the range [0,1].

Synthetic Task Generation. We adapt the image generation procedure involving random synthetic shapes described in SynthMorph [37] to produce a thousand synthetic tasks to be used alongside the medical tasks during training. We detail the generation process and include examples of synthetic tasks in supplemental Section D.

5. Experiments

We start by describing experimental details. The first set of experiments compares the performance of UniverSeg in the held-out datasets against several single-pass methods used in few-shot learning. We then report on a variety of analyses, including ablations of modeling decisions, and the effect of training task diversity, support set size, and number of examples available for a new task.

5.1. Experimental Setup

Model. We implement the network in UniverSeg (Figure 3) using an encoder with 5 CrossBlock stages and a decoder with 4 stages, with 64 output features per stage and LeakyReLU non-linearities after each convolution. We use bilinear interpolation when downsampling or upsampling.

Data. For each dataset d, we construct three disjoint splits $d = \{d_{\text{support}}, d_{\text{dev}}, d_{\text{test}}\}$ with 60%, 20%, and 20% of the subjects, respectively. Similar to dataset generalization [103], we divide the available datasets into a training set $\mathcal{D}^T$ and a held-out test set $\mathcal{D}^H$. We train models using the support and development splits of the training datasets $\{d_{\text{support}} | d \in \mathcal{D}^T\}$. We performed model selection and hyper-parameter tuning using the development split of held-out dataset WBC, and trained models until they stopped improving in the d_{dev} split, averaged across the held-out datasets. We report results using the unseen test split of the held-out datasets $\{d_{\text{test}} | d \in \mathcal{D}^H\}$. Support set image-label pairs are sampled with replacement from each dataset's support split.

For held-out datasets, we evaluated three datasets containing anatomies represented in the training datasets (ACDC [10] and SCD [85] (heart), and STARE[40] (retinal blood vessels)), and three datasets of anatomies not covered by the rest of MegaMedical (PanDental [2] (mandible), SpineWeb [114] (vertebrae), and WBC [115] (white blood cells).

Few-Shot Baselines. We compare UniverSeg models to three segmentation methods from the few-shot (FS) literature, since these approaches also predict the segmentation of a query image given a support set of image-label pairs, although they were designed for the low-data regime. SE-net [88] features a fully-convolutional network, squeeze-excitation blocks, and a UNet-like model architecture. ALP-Net [80] and PANet [105], employ prototypical networks that extract prototypes from their inputs to match the given query with the support set. While ALPNet also employs a self-supervised method to generate additional label maps in settings with few tasks, we omit this step since MegaMedical includes a large collection of tasks.

Unlike UniverSeg, these methods were designed to generalize to similar tasks, such as different labels in the same anatomy and image type, or different modalities for the same anatomy. To make the comparison to UniverSeg fair, we make several additions to the training and inference procedures of these baselines as described below, and chose the best performing variant of each baseline.

Supervised Task-Specific Models. While it is often impractical for clinical researchers to train individual networks for each task, for evaluation we train a set of task-specific networks to serve as an upper bound of supervised performance on the held-out datasets. We employ the widely-used nnUNet [42], which automatically configures the model and training pipeline based on data properties. Each model is task-specific, using the support and development splits for training and model selection, respectively. We report results on the test split.

Model	#Params	Runtime ms	Dice Score
PANet	14.71	240.0 ± 1.8	41.8 ± 1.3
ALPNet	43.02	527.7 ± 8.7	47.8 ± 1.1
SENet	0.92	4.1 ± 0.8	50.1 ± 1.3
UniverSeg (ours)	1.18	142.0 ± 0.4	$\mathbf{71.8 \pm 0.9}$
nnUNet (sup.)	$17\times$ 1.87	$17\times 1.4{\cdot}10^7$	84.4 ± 1.0

Table 1: **Performance Summary**. For UniverSeg and each FS baseline we report model size (in millions), inference runtime, and average held-out Dice score (with bootstrapping standard deviation) . As an upper bound, we include the set of 17 individually trained task-specific nnUNets for the 6 held-out datasets, where their run-time is their cumulative required training time.

Evaluation. We evaluate models on the held-out datasets $\mathcal{D}^H$ using the test split for query images and the support split for support-sets. For all methods, unless specified otherwise, we perform 5 independent predictions per test subject using randomly drawn support sets, and ensemble the predictions. We enforce that the same random support sets are used for all methods. We evaluate predictions using the Dice score [21] (0 - 100, 0=no overlap, 100=perfect match), which quantifies the region overlap between two regions and is widely used in medical segmentation. For tasks with more than one label, we average Dice across all labels. For datasets with multiple tasks, we average performance across all tasks. We estimate prediction variability using subject bootstrapping, with 1,000 independent repetitions. At each repetition, we treat each task independently, sampling subjects with replacement, and report the standard deviation across bootstrapped estimates.

Training. We train networks with the Adam optimizer [53] and soft Dice loss [74, 100]. For the ALPNet and PANet baselines, we add a prototypical loss term as described in their original works. Models trained with cross-entropy performed substantially worse than soft Dice.

While the original baseline methods were not introduced with significant data augmentation, we trained all UniverSeg and FS models with and without the proposed augmentation transformations, and report results on the best-performing setting. Unless specified otherwise, models are trained using a support size of 64. While the baselines were originally designed with small support sizes (1 or 5) as they tackled the few-shot setting, we found that training and evaluating them with larger support sizes improved their performance.

Implementation. We provide additional implementation and experimental details in supplemental Section B. Code and pre-trained model weights for UniverSeg are available at `https://universeg.csail.mit.edu`.

5.2. Task Generalization Results

First, we compare the segmentation quality of UniverSeg with FS baselines and the task-specific upper bounds. Our primary goal is to assess the effectiveness of UniverSeg in solving tasks from unseen datasets. Figure 4 presents the average Dice scores per dataset for each method, and Figure 5 presents example segmentation results for each method and dataset.

Few-shot methods. UniverSeg significantly outperforms all FS methods in all held-out datasets. For each FS method, we report the best-performing model, which involved adding components of the UniverSeg training pipeline. In the supplemental material, we show that few-shot methods perform worse when trained with a support set size of 1 and without ensembling, as they were originally introduced.

UniverSeg outperforms the highest performing baseline for all datasets with Dice improvements ranging from 7.3 to 34.9. Figure 5 also shows clear qualitative improvements in the predicted segmentations. Given the similarities between SENet and UniverSeg (fully convolutional UNet-like structure), these results suggest that the proposed CrossBlock is better suited to transferring spatial information from the support set to the query. Table 1 shows that UniverSeg also requires fewer model parameters than PANet, ALPNet, and the nnUNets, and a similar number to SENet.

Task-specific networks. For some datasets like PanDental or WBC, UniverSeg performs competitively with the supervised task-specific networks, which were extensively trained on each of the held-out tasks, and are unfeasible to run in many clinical research settings. Moreover, from the qualitative results of Figure 5, we observe that segmentations produced by UniverSeg more closely match those of the supervised baselines than those of any other few-shot segmentation task, especially in challenging datasets like SpineWeb or STARE.

5.3. Analysis

We analyze how several of the data, model, and training decisions affect the performance of UniverSeg.

Task Quantity and Diversity. We study the effect of the number of datasets and individual tasks used for training UniverSeg. We leave out synthetic tasks for this experiment, and train models on random subsets of the MegaMedical training datasets.

Figure 6 presents performance on the held-out datasets for different random subsets of training datasets. We find that having more training tasks improves the performance on held-out tasks. In some scenarios, the *choice* of datasets has a substantial effect. For instance, for models trained with 10% of the datasets, the best model outperforms the worst one by 17.3 Dice points, and comparing those subsets

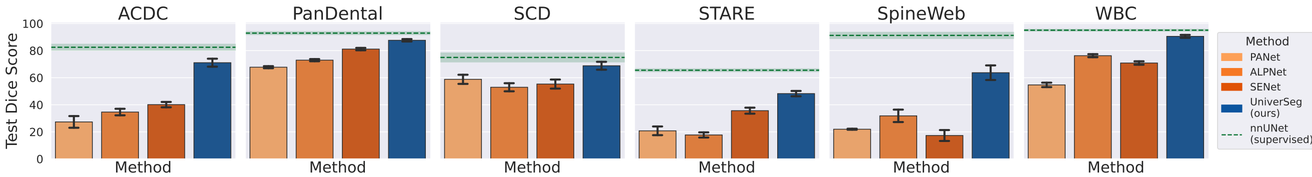

Figure 4: **Average Dice score per each held out dataset**. Performance of UniverSeg and several few-shot baselines, and the upper bound of each dataset determined by the individual fully-trained networks. For each of the unseen datasets, we average across tasks and subjects, and show the bootstrap variability in the error bars.

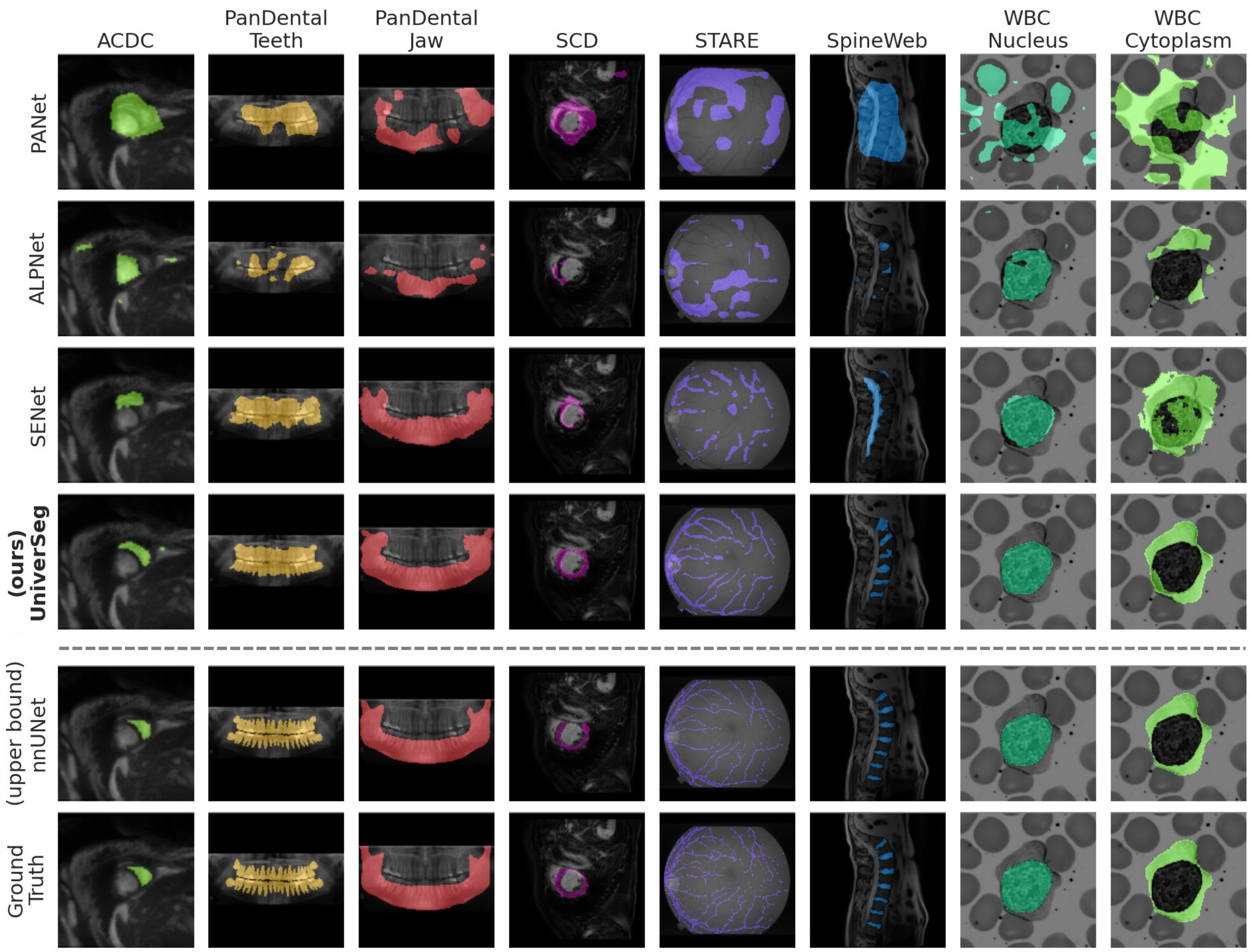

Figure 5: **Example model predictions for unseen tasks**. For a randomly sampled image per held-out task, we visualize the predictions of UniverSeg, few-shot baselines, and individually trained nnUNet models, along with ground truth maps.

we find that the best performing one was trained on a broad set of anatomies including heart, abdomen, brain, and eyes; while the least accurate model was trained on less common lesion tasks, leading to worse generalization.

Ablation of Training Strategies. We perform an ablation study over the three main techniques we employ for increasing data and task diversity during training: in-task augmentation, task augmentation, and synthetic tasks.

Table 2 shows that all proposed strategies lead to improvements in model performance, with the best results achieved when using all strategies jointly, providing a boost of 9 Dice points over no augmentations or synthetic tasks. Incorporating task augmentation leads to the largest individual improvement of 7.7 Dice points. Remarkably, the model trained using only synthetic data performs surprisingly well on the medical held-out tasks despite having never been exposed to medical training data. These results suggest that increasing image and task diversity during training, even artificially, has

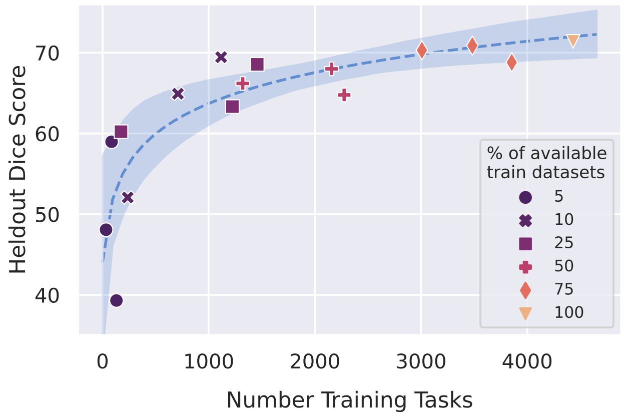

Figure 6: **Average held-out Dice versus the number of training tasks**. Points represent individual UniverSeg networks trained on a percentage of available training datasets and shown in terms of the number of underlying training tasks. In blue, we report a logarithmic fit to the data and 95% confidence intervals obtained by bootstrapped fits.

Synth	Medical	In-Task	Task	Dice Score
✓				61.7 ± 1.5
	✓			62.7 ± 1.1
✓	✓			64.5 ± 1.0
	✓	✓		67.0 ± 0.9
	✓		✓	70.4 ± 1.3
	✓	✓	✓	70.0 ± 1.5
✓	✓	✓	✓	**71.8 ± 0.9**

Table 2: **Training Strategies Ablation**. Average held-out Dice for UniverSeg models trained with different combinations of proposed techniques to increase task diversity: in-task augmentation, task augmentation, and synthetic tasks.

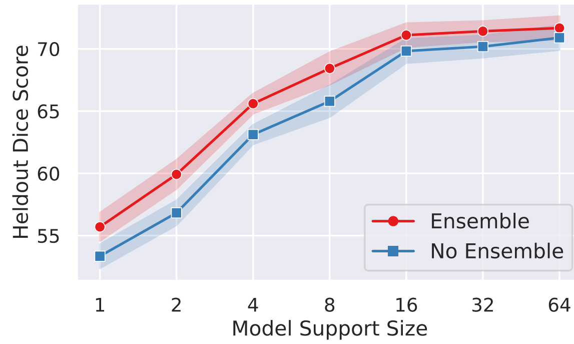

Figure 7: **Effect of support size.** Relationship between models trained at certain support sizes and their average held-out Dice score. Results improve with higher support size, with ensembling consistently helping.

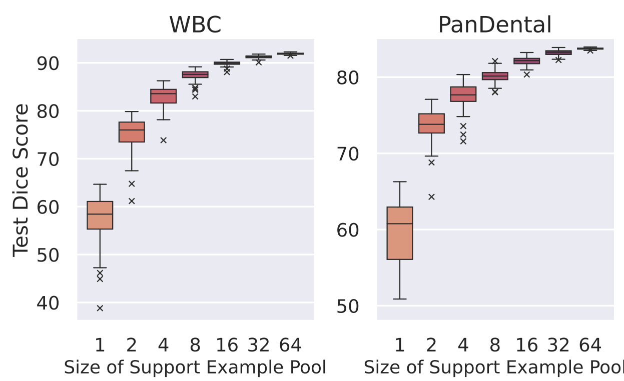

Figure 8: **Effect of available data at inference.** UniverSeg predictions using a limited d_{support} example pool on the held-out WBC and PanDental datasets. For each size, we perform 100 repetitions using different random subsets.

a substantial effect on how the model generalizes to unseen segmentation tasks.

Support Set Size. We study the effect of support size on models trained with support sizes N from 1 to 64.

Figure 7 shows that the best results are achieved with large training support set sizes, with the average held-out Dice rapidly improving from 53.7 to 69.9 for supports sizes from 1 to 16, and then providing diminishing returns at greater support sizes, with a maximum of 71 Dice at support size 64. We find that ensembling predictions leads to consistent improvements in all cases, with greater improvements of 2.4-3.1 Dice points for small support sets ($N < 16$).

Limited Example Data. Since manually annotating examples from new tasks is expensive for medical data, we investigate how the number of labeled images affects the performance of UniverSeg. We study UniverSeg when using a limited amount of labeled examples N at inference, for $N = 1, 2, \ldots, 64$. We perform 100 repetitions for each size, each corresponding to an independent random subset of the data. Here, the support set contains all available data for inference, and thus we do not perform ensembling.

Figure 8 presents results for the WBC and PanDental held-out datasets, which have 108 and 116 examples in their d_{support} splits respectively. For small values of support size N, we observe a large variance caused by very diverse support sets. As N increases, we observe that average segmentation quality monotonically improves and the variance from the sample of available data examples is greatly reduced.

Support Set Ensembling. We study the effect of varying the support size N at inference, and number K of predictions being ensembled. We first sample 100 independent support sets for each inference support size N. Then, for each ensembling amount K, we compute ensembled predictions by averaging K independently drawn predictions.

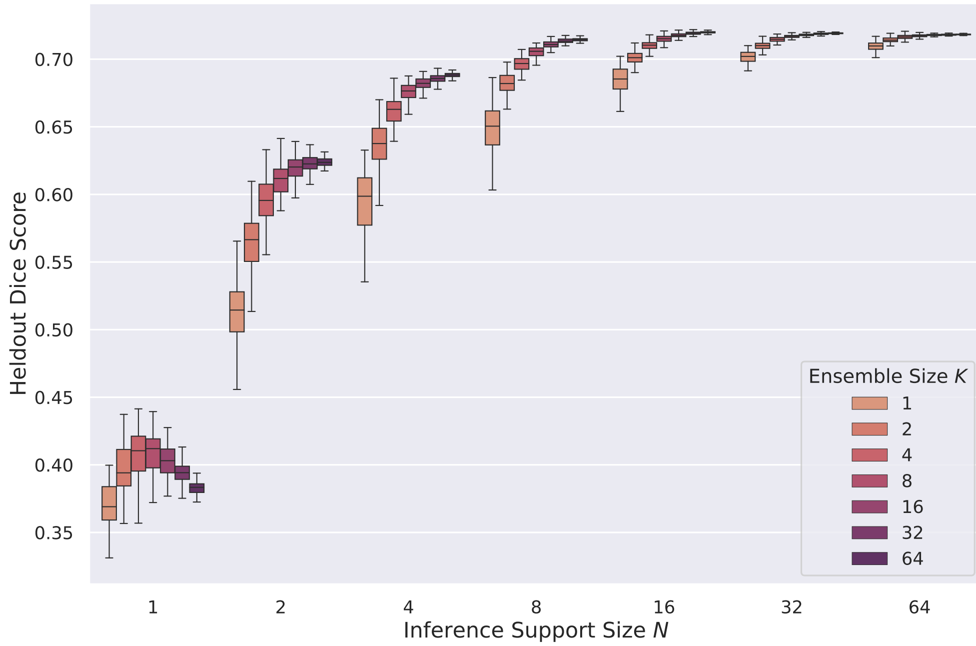

Figure 9: **Ensembling predictions at different inference support sizes.** Average held-out test Dice Score for different settings of ensembling and support size. For each inference support size N, we report the results (in average held-out Dice Score) of taking 100 predictions ($K = 1$) and ensembling by averaging in groups of size K, performing 100 repetitions for each K. The value boxes report quantiles over the 100 values for each setting and find that increasing either K or N leads to improved model performance, with N having a significantly larger effect than K.

Figure 9 shows that given a certain support size, increasing the ensemble size leads to monotonic improvements and reduced variance, likely by being less dependent on the specific examples in the support set. The performance also monotonically improves with increased support size N, which has a significantly larger effect on segmentation accuracy than increasing the ensemble size. For instance, non-ensembled predictions with support size 64 ($N = 64$, $K = 1$) are better than heavily ensembled predictions with smaller support sizes ($N = 2, 4, 8$ and $K = 64$), even though the latter uses more support examples. This suggests that UniverSeg models exploit information coming from the support examples in a fundamentally different way than existing ensembling techniques used in FS learning.

6. Discussion and Conclusion

We introduce UniverSeg, an approach for learning a *single* task-agnostic model for medical image segmentation. We use a large and diverse collection of open-access medical segmentation datasets to train UniverSeg, which is capable of generalizing to unseen anatomies and tasks. UniverSeg introduces the idea that segmentation tasks from diverse biomedical domains can be defined, or prompted, by a set of segmentation examples. We introduce a novel *cross-convolution* operation that interacts the query and support representations at different scales.

In our experiments, UniverSeg substantially outperforms existing few-shot methods in all held-out datasets. Through extensive ablation studies, we conclude that UniverSeg performance is strongly dependent on task diversity during training and support set diversity during inference. This highlights the utility of UniverSeg facilitating variably-sized support sets, enabling flexibility to potential users' datasets.

Limitations. In this work, we focused on demonstrating and thoroughly analyzing the core idea of UniverSeg, using 2D data and single labels. We are excited by future extensions to segment 3D volumes using 2.5D or 3D models and multi-label maps, and further closing the gap with the upper bounds.

Outlook. UniverSeg promises to easily adapt to new segmentation tasks determined by scientists and clinical researchers, without model retraining which is often impractical for them.

Acknowledgements

We thank Aniruddh Raghu for helpful discussion and feedback. Victor Butoi is supported by the NSF GRFP Fellowship, and Jose Javier is supported by Quanta Computer, Inc, and the project was supported by NIH R01AG053949, R01AG064027, and NSF CAREER 1748377.

References

[1] Thyroid ultrasound cine-clip, Oct 2021. 16

[2] Amir Hossein Abdi, Shohreh Kasaei, and Mojdeh Mehdizadeh. Automatic segmentation of mandible in panoramic x-ray. *Journal of Medical Imaging*, 2(4):044003, 2015. 5, 16

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 3

[4] Laith Alzubaidi, Muthana Al-Amidie, Ahmed Al-Asadi, Amjad J. Humaidi, Omran Al-Shamma, Mohammed A. Fadhel, Jinglan Zhang, J. Santamaría, and Ye Duan. Novel transfer learning approach for medical imaging with limited labeled data. *Cancers*, 13(7), 2021. 3

[5] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021. 5, 16

[6] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017. 16

[7] Guha Balakrishnan, Amy Zhao, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging*, 38(8):1788–1800, 2019. 15

[8] Ivana Balažević, David Steiner, Nikhil Parthasarathy, Relja Arandjelović, and Olivier J Hénaff. Towards in-context scene understanding. *arXiv preprint arXiv:2306.01667*, 2023. 3

[9] Sophia Bano, Francisco Vasconcelos, Luke M Shepherd, Emmanuel Vander Poorten, Tom Vercauteren, Sebastien Ourselin, Anna L David, Jan Deprest, and Danail Stoyanov. Deep placental vessel segmentation for fetoscopic mosaicking. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, pages 763–773. Springer, 2020. 16

[10] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018. 5, 16

[11] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019. 5, 16

[12] Benjamin Billot, Douglas Greve, Koen Van Leemput, Bruce Fischl, Juan Eugenio Iglesias, and Adrian V Dalca. A learning strategy for contrast-agnostic mri segmentation. *arXiv preprint arXiv:2003.01995*, 2020. 20

[13] Nicholas Bloch, Anant Madabhushi, Henkjan Huisman, John Freymann, Justin Kirby, Michael Grauer, Andinet Enquobahrie, Carl Jaffe, Larry Clarke, and Keyvan Farahani. Nci-isbi 2013 challenge: automated segmentation of prostate structures. *The Cancer Imaging Archive*, 370(6):5, 2015. 5, 16

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

[15] Mateusz Buda, Ashirbani Saha, and Maciej A Mazurowski. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in biology and medicine*, 109:218–225, 2019. 16

[16] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997. 2

[17] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 1

[18] Adrian V Dalca, John Guttag, and Mert R Sabuncu. Anatomical priors in convolutional networks for unsupervised biomedical segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9290–9299, 2018. 16

[19] Etienne Decenciere, Guy Cazuguel, Xiwei Zhang, Guillaume Thibault, J-C Klein, Fernand Meyer, Beatriz Marcotegui, Gwénolé Quellec, Mathieu Lamard, Ronan Danno, et al. Teleophta: Machine learning and image processing methods for teleophthalmology. *Irbm*, 34(2):196–203, 2013. 16

[20] Aysen Degerli, Morteza Zabihi, Serkan Kiranyaz, Tahir Hamid, Rashid Mazhar, Ridha Hamila, and Moncef Gabbouj. Early detection of myocardial infarction in low-quality echocardiography. *IEEE Access*, 9:34442–34453, 2021. 16

[21] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. 6, 17

[22] Hao Ding, Changchang Sun, Hao Tang, Dawen Cai, and Yan Yan. Few-shot medical image segmentation with cycle-resemblance attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2488–2497, 2023. 3

[23] Jose Dolz, Karthik Gopinath, Jing Yuan, Herve Lombaert, Christian Desrosiers, and Ismail Ben Ayed. Hyperdense-net: A hyper-densely connected cnn for multi-modal image segmentation. *IEEE Transactions on Medical Imaging*, 38(5):1116–1126, 2019. 1, 2

[24] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117, 2004. 2

[25] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 3

[26] Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012. 15

[27] Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles RG Guttmann, Frank-Erik de Leeuw, Clare M Tempany, Bram van Ginneken, et al. Transfer learning for domain adaptation in mri: Application in brain lesion segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 516–524. Springer, 2017. 3

[28] Randy L Gollub, Jody M Shoemaker, Margaret D King, Tonya White, Stefan Ehrlich, Scott R Sponheim, Vincent P Clark, Jessica A Turner, Bryon A Mueller, Vince Magnotta, et al. The mcic collection: a shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia. *Neuroinformatics*, 11:367–388, 2013. 5, 15, 16

[29] Ioannis S Gousias, A David Edwards, Mary A Rutherford, Serena J Counsell, Jo V Hajnal, Daniel Rueckert, and Alexander Hammers. Magnetic resonance imaging of the newborn brain: manual segmentation of labelled atlases in term-born and preterm infants. *Neuroimage*, 62(3):1499–1509, 2012. 16

[30] Ioannis S Gousias, Daniel Rueckert, Rolf A Heckemann, Leigh E Dyet, James P Boardman, A David Edwards, and Alexander Hammers. Automatic segmentation of brain mris of 2-year-olds into 83 regions of interest. *Neuroimage*, 40(2):672–684, 2008. 16

[31] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical Image Analysis*, 58:101563, 2019. 16

[32] Stine Hansen, Srishti Gautam, Robert Jenssen, and Michael Kampffmeyer. Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels. *Medical Image Analysis*, 78:102385, 2022. 3

[33] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4918–4927, 2019. 3

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 17

[35] Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, page 101821, 2020. 5, 16

[36] Moritz R Hernandez Petzsche, Ezequiel de la Rosa, Uta Hanning, Roland Wiest, Waldo Valenzuela, Mauricio Reyes, Maria Meyer, Sook-Lei Liew, Florian Kofler, Ivan Ezhov, et al. Isles 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific data*, 9(1):762, 2022. 5, 16

[37] Malte Hoffmann, Benjamin Billot, Douglas N Greve, Juan Eugenio Iglesias, Bruce Fischl, and Adrian V Dalca. Synthmorph: learning contrast-invariant registration without acquired images. *IEEE Transactions on Medical Imaging*, 41(3):543–558, 2022. 5, 20

[38] Andrew Hoopes, Malte Hoffmann, Douglas N. Greve, Bruce Fischl, John Guttag, and Adrian V. Dalca. Learning the effect of registration hyperparameters with hypermorph. volume 1, pages 1–30, 2022. 15, 16

[39] Andrew Hoopes, Jocelyn S Mora, Adrian V Dalca, Bruce Fischl, and Malte Hoffmann. Synthstrip: skull-stripping for any brain image. *NeuroImage*, 260:119474, 2022. 20

[40] AD Hoover, Valentina Kouznetsova, and Michael Goldbaum. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical imaging*, 19(3):203–210, 2000. 5, 16

[41] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. 1, 2

[42] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 1, 2, 5

[43] Yuanfeng Ji, Haotian Bai, Jie Yang, Chongjian Ge, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *arXiv preprint arXiv:2206.08023*, 2022. 5, 16

[44] Zhexin Jiang, Hao Zhang, Yi Wang, and Seok-Bum Ko. Retinal blood vessel segmentation using fully convolutional network with transfer learning. *Computerized Medical Imaging and Graphics*, 68:1–15, 2018. 3

[45] Konstantinos Kamnitsas, Enzo Ferrante, Sarah Parisot, Christian Ledig, Aditya V Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. Deepmedic for brain tumor segmentation. In *International workshop on Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries*, pages 138–149. Springer, 2016. 1

[46] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017. 2

[47] Rashed Karim, R James Housden, Mayuragoban Balasubramaniam, Zhong Chen, Daniel Perry, Ayesha Uddin, Yosra Al-Beyatti, Ebrahim Palkhi, Prince Acheampong, Samantha Obom, et al. Evaluation of current algorithms for segmentation of scar tissue from late gadolinium enhancement cardiovascular magnetic resonance of the left atrium: an open-access grand challenge. *Journal of Cardiovascular Magnetic Resonance*, 15(1):1–17, 2013. 16

[48] A. Emre Kavur, N. Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, Bora Baydar, Dmitry Lachinov, Shuo Han, Josef Pauli, Fabian Isensee, Matthias Perkonigg, Rachana Sathish, Ronnie Rajan, Debdoot Sheet, Gurbandurdy Dovletov, Oliver Speck, Andreas Nürnberger, Klaus H. Maier-Hein, Gözde Bozdağı Akar, Gözde Ünal, Oğuz Dicle, and M. Alper Selver. CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, Apr. 2021. 16

[49] Ali Emre Kavur, M. Alper Selver, Oğuz Dicle, Mustafa Barış, and N. Sinem Gezer. CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data. Apr. 2019. 5, 15

[50] Ali Emre Kavur, M. Alper Selver, Oğuz Dicle, Mustafa Barış, and N. Sinem Gezer. CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data, Apr. 2019. 16

[51] Rabindra Khadka, Debesh Jha, Steven Hicks, Vajira Thambawita, Michael A Riegler, Sharib Ali, and Pål Halvorsen. Meta-learning with implicit gradients in a few-shot setting for medical image segmentation. *Computers in Biology and Medicine*, 143:105227, 2022. 3

[52] Pulkit Khandelwal and Paul Yushkevich. Domain generalizer: A few-shot meta learning framework for domain generalization in medical imaging. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, pages 73–84. Springer, 2020. 3

[53] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6, 17

[54] Serkan Kiranyaz, Aysen Degerli, Tahir Hamid, Rashid Mazhar, Rayyan El Fadil Ahmed, Rayaan Abouhasera, Morteza Zabihi, Junaid Malik, Ridha Hamila, and Moncef Gabbouj. Left ventricular wall motion estimation by active polynomials for acute myocardial infarction detection. *IEEE Access*, 8:210301–210317, 2020. 16

[55] Hugo J Kuijf, J Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adria Casamitjana, et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE transactions on medical imaging*, 38(11):2556–2568, 2019. 5, 16

[56] Maria Kuklisova-Murgasova, Paul Aljabar, Latha Srinivasan, Serena J Counsell, Valentina Doria, Ahmed Serag, Ioannis S Gousias, James P Boardman, Mary A Rutherford, A David Edwards, et al. A dynamic 4d probabilistic atlas of the developing brain. *NeuroImage*, 54(4):2750–2763, 2011. 5, 16

[57] Zoé Lambert, Caroline Petitjean, Bernard Dubray, and Su Kuan. Segthor: segmentation of thoracic organs at risk in ct images. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. IEEE, 2020. 5, 16

[58] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, page 12, 2015. 5, 16

[59] Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019. 16

[60] Guillaume Lemaître, Robert Martí, Jordi Freixenet, Joan C Vilanova, Paul M Walker, and Fabrice Meriaudeau. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review. *Computers in biology and medicine*, 60:8–31, 2015. 5, 16

[61] Mingchao Li, Yuhan Zhang, Zexuan Ji, Keren Xie, Songtao Yuan, Qinghuai Liu, and Qiang Chen. Ipn-v2 and octa-500: Methodology and dataset for retinal image segmentation. *arXiv preprint arXiv:2012.07261*, 2020. 5, 16

[62] Yiwen Li, Yunguan Fu, Iani Gayo, Qianye Yang, Zhe Min, Shaheer Saeed, Wen Yan, Yipei Wang, J Alison Noble, Mark Emberton, et al. Prototypical few-shot segmentation for cross-institution male pelvic structures with spatial registration. *arXiv preprint arXiv:2209.05160*, 2022. 3

[63] Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Medical image analysis*, 18(2):359–373, 2014. 5, 16

[64] Quande Liu, Qi Dou, Lequan Yu, and Pheng Ann Heng. Ms-net: Multi-site network for improving prostate segmentation with heterogeneous mri data. *IEEE Transactions on Medical Imaging*, 2020. 2

[65] Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature methods*, 9(7):637–637, 2012. 16

[66] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. *Advances in neural information processing systems*, 29, 2016. 2

[67] Xiangde Luo, Wenjun Liao, Jianghong Xiao, Tao Song, Xiaofan Zhang, Kang Li, Guotai Wang, and Shaoting Zhang. Word: Revisiting organs segmentation in the whole abdominal region. *arXiv preprint arXiv:2111.02403*, 2021. 5, 16

[68] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, Shucheng Cao, Qi Zhang, Shangqing Liu, Yunpeng Wang, Yuhui Li, Jian He, and Xiaoping Yang. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 5, 16

[69] Yuhui Ma, Huaying Hao, Jianyang Xie, Huazhu Fu, Jiong Zhang, Jianlong Yang, Zhen Wang, Jiang Liu, Yalin Zheng, and Yitian Zhao. Rose: a retinal oct-angiography vessel segmentation dataset and new model. *IEEE Transactions on Medical Imaging*, 40(3):928–939, 2021. 5, 16

[70] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007. 5, 15, 16

[71] Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, et al. The parkinson progression marker initiative (ppmi). *Progress in neurobiology*, 95(4):629–635, 2011. 5, 15, 16

[72] Maciej A Mazurowski, Kal Clark, Nicholas M Czarnek, Parisa Shamsesfandabadi, Katherine B Peters, and Ashirbani Saha. Radiogenomics of lower-grade glioma: algorithmically-assessed tumor shape is associated with tumor genomic subtypes and patient outcomes in a multi-institutional study with the cancer genome atlas data. *Journal of neuro-oncology*, 133:27–35, 2017. 5, 16

[73] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 16

[74] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 1, 4, 6

[75] Pim Moeskops, Jelmer M Wolterink, Bas HM van der Velden, Kenneth GA Gilhuijs, Tim Leiner, Max A Viergever, and Ivana Išgum. Deep learning for multi-task medical image segmentation in multiple modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 478–486. Springer, 2016. 2

[76] Fernando Navarro, Suprosanna Shit, Ivan Ezhov, Johannes Paetzold, Andrei Gafita, Jan C Peeken, Stephanie E Combs, and Bjoern H Menze. Shape-aware complementary-task learning for multi-organ segmentation. In *International Workshop on Machine Learning in Medical Imaging*, pages 620–627. Springer, 2019. 2

[77] Khoi Nguyen and Sinisa Todorovic. Feature weighting and boosting for few-shot segmentation. *CoRR*, abs/1909.13140, 2019. 3

[78] Alex Nichol and John Schulman. Reptile: a scalable metalearning algorithm. *arXiv preprint arXiv:1803.02999*, 2(3):4, 2018. 3

[79] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022. 3

[80] Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In *European Conference on Computer Vision*, pages 762–780. Springer, 2020. 3, 5

[81] Prashant Pandey, Mustafa Chasmai, Tanuj Sur, and Brejesh Lall. Robust prototypical few-shot organ segmentation with regularized neural-odes. *arXiv preprint arXiv:2208.12428*, 2022. 3

[82] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 17

[83] Kelly Payette, Priscille de Dumast, Hamza Kebiri, Ivan Ezhov, Johannes C Paetzold, Suprosanna Shit, Asim Iqbal, Romesa Khan, Raimund Kottke, Patrice Grehten, et al. An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. *Scientific Data*, 8(1):1–14, 2021. 16

[84] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid), 2018. 5, 16

[85] Perry Radau, Yingli Lu, Kim Connelly, Gideon Paul, AJWG Dick, and Graham Wright. Evaluation framework for algorithms segmenting short axis cardiac mri. *The MIDAS Journal-Cardiac MR Left Ventricle Segmentation Challenge*, 49, 2009. 5, 16

[86] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32, 2019. 2, 3

[87] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1, 2

[88] Abhijit Guha Roy, Shayan Siddiqui, Sebastian Pölsterl, Nassir Navab, and Christian Wachinger. 'squeeze & excite'guided few-shot segmentation of volumetric images. *Medical image analysis*, 59:101587, 2020. 3, 5

[89] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, SQ Truong, CD Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, AY Ng, et al. Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation. *MedRxiv*, 2021. 5, 16

[90] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018. 2

[91] Jun Seo, Young-Hyun Park, Sung Whan Yoon, and Jaekyun Moon. Task-adaptive feature transformer with semantic enrichment for few-shot segmentation. *arXiv preprint arXiv:2202.06498*, 2022. 3

[92] Ahmed Serag, Paul Aljabar, Gareth Ball, Serena J Counsell, James P Boardman, Mary A Rutherford, A David Edwards, Joseph V Hajnal, and Daniel Rueckert. Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression. *Neuroimage*, 59(3):2255–2265, 2012. 16

[93] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, Moira SN Berens, Cas Van Den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Evelina Fantacci, Bram Geurts, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. *Medical image analysis*, 42:1–13, 2017. 5, 16

[94] Neeraj Sharma and Lalit M Aggarwal. Automated medical image segmentation techniques. *Journal of medical physics/Association of Medical Physicists of India*, 35(1):3, 2010. 1

[95] Qianqian Shen, Yanan Li, Jiyong Jin, and Bin Liu. Q-net: Query-informed few-shot medical image segmentation. *arXiv preprint arXiv:2208.11451*, 2022. 3

[96] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019. 5, 16

[97] Rishav Singh, Vandana Bharti, Vishal Purohit, Abhinav Kumar, Amit Kumar Singh, and Sanjay Kumar Singh. Metamed: Few-shot medical image classification using gradient-based meta-learning. *Pattern Recognition*, 120:108111, 2021. 3

[98] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 3

[99] Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4):501–509, 2004. 5, 15, 16

[100] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017. 6

[101] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016. 2

[102] Hao Tang, Xingwei Liu, Shanlin Sun, Xiangyi Yan, and Xiaohui Xie. Recurrent mask refinement for few-shot medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3918–3928, 2021. 3

[103] Eleni Triantafillou, Hugo Larochelle, Richard Zemel, and Vincent Dumoulin. Learning a universal template for few-shot dataset generalization. In *International Conference on Machine Learning*, pages 10424–10433. PMLR, 2021. 5

[104] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016. 3

[105] Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. *CoRR*, abs/1908.06391, 2019. 3, 5

[106] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839, 2023. 3

[107] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023. 3

[108] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021. 3

[109] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5217–5226, 2019. 3

[110] Penghao Zhang, Jiayue Li, Yining Wang, and Judong Pan. Domain adaptation for medical image segmentation: a meta-learning method. *Journal of Imaging*, 7(2):31, 2021. 3

[111] Yingtao Zhang, Min Xian, Heng-Da Cheng, Bryar Shareef, Jianrui Ding, Fei Xu, Kuan Huang, Boyu Zhang, Chunping Ning, and Ying Wang. Busis: A benchmark for breast ultrasound image segmentation. In *Healthcare*, volume 10, page 729. MDPI, 2022. 16

[112] Yao Zhang, Jiawei Yang, Jiang Tian, Zhongchao Shi, Cheng Zhong, Yang Zhang, and Zhiqiang He. Modality-aware mutual learning for multi-modal medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 589–599. Springer, 2021. 3

[113] Amy Zhao, Guha Balakrishnan, Fredo Durand, John V Guttag, and Adrian V Dalca. Data augmentation using learned transformations for one-shot medical image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8543–8553, 2019. 3

[114] Guoyan Zheng, Chengwen Chu, Daniel L Belavỳ, Bulat Ibragimov, Robert Korez, Tomaž Vrtovec, Hugo Hutt, Richard Everson, Judith Meakin, Isabel Lŏpez Andrade, et al. Evaluation and comparison of 3d intervertebral disc localization and segmentation methods for 3d t2 mr data: A grand challenge. *Medical image analysis*, 35:327–344, 2017. 5, 15, 16

[115] Xin Zheng, Yong Wang, Guoyou Wang, and Jianguo Liu. Fast and robust segmentation of white blood cell images by self-supervised learning. *Micron*, 107:55–71, 2018. 5, 15, 16

[116] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021. 3