This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Show and Segment: Universal Medical Image Segmentation via In-Context Learning

Yunhe Gao[2], Di Liu[2], Zhuowei Li[2], Yunsheng Li[1], Dongdong Chen[1], Mu Zhou[2], Dimitris N. Metaxas[2]

[1]Microsoft GenAI [2]Rutgers University

Abstract

*Medical image segmentation remains challenging due to the vast diversity of anatomical structures, imaging modalities, and segmentation tasks. While deep learning has made significant advances, current approaches struggle to generalize as they require task-specific training or fine-tuning on unseen classes. We present **Iris**, a novel **I**n-context **R**eference **I**mage guided **S**egmentation framework that enables flexible adaptation to novel tasks through the use of reference examples without fine-tuning. At its core, Iris features a lightweight context task encoding module that distills task-specific information from reference context image-label pairs. This rich context embedding information is used to guide the segmentation of target objects. By decoupling task encoding from inference, Iris supports diverse strategies from one-shot inference and context example ensemble to object-level context example retrieval and in-context tuning. Through comprehensive evaluation across twelve datasets, we demonstrate that Iris performs strongly compared to task-specific models on in-distribution tasks. On seven held-out datasets, Iris shows superior generalization to out-of-distribution data and unseen classes. Further, Iris's task encoding module can automatically discover anatomical relationships across datasets and modalities, offering insights into medical objects without explicit anatomical supervision.*

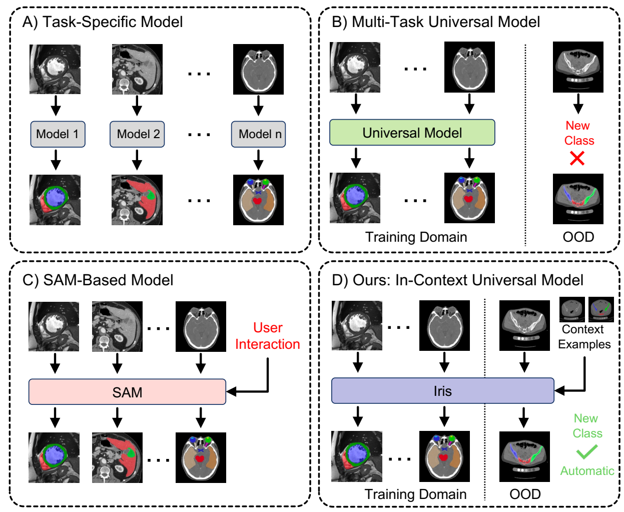

Figure 1. Comparison of medical image segmentation approaches. A) Task-specific models require training separate models for each task, limiting their flexibility and scalability. B) Multi-task universal models can handle diverse tasks and imaging modalities, but fail on novel classes. C) SAM-based foundation models enable flexible segmentation through user interactions, but impractical for high-throughput automated processing. D) Our proposed Iris combines automatic processing with flexible adaptation via in-context learning, enabling both seen and unseen task segmentation without any manual interaction or retraining.

1. Introduction

The accurate segmentation of anatomical structures in medical images is fundamental for clinical practice and biomedical research, enabling precise diagnosis [11, 41] and treatment planning [37]. While deep learning has demonstrated remarkable success [32, 46], the vast diversity of anatomical structures, imaging modalities, and clinical tasks poses long-standing challenges for developing truly generalizable solutions. Current efforts typically focus on disease-specific tasks or a limited set of anatomical structures [7, 13, 14, 17, 20, 27–29, 59], struggling to handle the heterogeneous landscape of medical imaging that spans diverse modalities, body regions, and diseases [38, 52].

These task-specific methods show critical limitations compared to human experts' capabilities. First, existing models often perform poorly on out-of-distribution examples [60]—a common scenario in medical imaging where variations arise from different imaging centers, patient populations, and acquisition protocols. Second, traditional segmentation models, while achieving a high accuracy on their trained tasks, lack the adaptability to handle novel classes without extensive retraining or fine-tuning [56]. This dilemma fundamentally limits task-specific models' applicability in dynamic clinical settings and research environments, where new segmentation tasks continue to emerge over the course of real-world practice.

Recent research has explored several directions to address these challenges (Figure 1). Universal medical segmentation models [30, 44, 50, 54] attempt to leverage synergy among multiple tasks across diverse datasets to learn robust representations, yet struggling with unseen classes and requiring fine-tuning. Foundation models with interactive capabilities, such as SAM [23] and its medical variants [9, 34, 45, 58], offer flexibility via user prompts. But they require multiple interactions for optimal segmentation results, especially for complex 3D structures, and lack the efficiency for large-scale automated analysis. In addition, in-context learning (ICL) methods [5, 39] show promise in automatically handling arbitrary new tasks through a few reference examples, but current methods exhibit suboptimal performance compared to task-specific models and suffer from computational inefficiencies, requiring expensive reference encoding during each inference step.

To address these fundamental challenges, we present Iris framework for universal medical image segmentation via in-context learning. At its core, Iris features a lightweight task encoding module that efficiently distills task-specific information from reference image-label pairs into compact task embeddings, which then guide the segmentation of target objects. Unlike existing ICL methods [5, 39], Iris decouples the task definition from query image inference, eliminating redundant context encoding while enabling flexible inference strategies, all coming with high computational efficiency.

Our main contributions include:

- A novel in-context learning framework for 3D medical images, enabling a strong adaptation to arbitrary new segmentation tasks without model retraining or fine-tuning.
- A lightweight task encoding module that captures task-specific information from reference examples, handling medical objects of varying sizes and shapes.
- Multiple flexible inference strategies suitable for different practical scenarios, including one-shot inference, context ensemble, object-level context retrieval, and in-context tuning.
- Comprehensive experiments on 19 datasets demonstrate Iris's superior performance across both in-distribution and challenging scenarios, particularly on held-out domains and novel anatomical structures. It extends to reveal the capability of automatically discovering meaningful anatomical relationships across datasets and modalities.

2. Related Work

Medical Universal Models. Universal medical image segmentation models aim to address the data heterogeneity across tasks and modalities while learning generalizable feature representations. Early works include multi-dataset learning through label conditioning [12], organ size constraints [61], and pseudo-label co-training [19]. Recent works are placed on sophisticated task encoding strategies. DoDNet [54] pioneered one-hot task vectors with an extension into TransDoDNet [49] using transformer backbones. Advances include CLIP-driven models using semantic encodings [30], task-specific heads in MultiTalent [44], and modality priors in Hermes [13]. UniSeg [50] introduced learnable task prompts and MedUniseg [51] unified 2D/3D image handling. Despite of substantial efforts, these universal models all require fine-tuning when assessing unseen classes. In contrast, Iris enables the segmentation of unseen classes only through a single reference image-label pair without any model finetuning.

SAM-based Interactive Models. Segment Anything Model (SAM) [23] emerges as a shifting paradigm of interactive segmentation via its prompt-based architecture and large-scale training. SAM's success has inspired a range of medical variants. Major examples include MedSAM [34] with 1.5M image-mask pairs for 2D segmentation, SAM-Med2D [9] trained on 4.6M images, and SAM-Med3D [45] extending to volumetric data with 22K 3D images. These efforts all require multiple prompts and interactive refinements, especially for analyzing complex objects in 3D scenarios. This interaction-dependent design becomes a bottleneck in high-throughput scenarios that an automated processing of large-scale datasets is much desired. As comparison, Iris addresses this limitation by defining tasks through context pairs, enabling fully automatic segmentation while maintaining a strong adaptability to new tasks.

Visual In-context Learning. In-context learning as introduced by GPT-3 [4] enables models to handle novel tasks through example-guided inference without a heavy retraining. In the vision community, Painter [47] and SegGPT [48] pioneered in-context segmentation through a mask image modeling framework. Alternative methods [33, 43, 55] explored SAM-based approaches through cross-image correspondence prompting, but their two-stage pipeline introduces redundant computation and heavily relies on SAM's capabilities, limiting their applicability to 3D medical images. Neuralizer [10] develop general tools on diverse neuroimaging tasks, like super-resolution denosing, etc. Recent works introduced specialized architectures for in-context segmentation [36]. For example, UniverSeg [5] is designed for in-context medical image segmentation, and Tyche [39] incorporated a stochastic inference. While these methods demonstrate promising capability on novel classes, they face two critical limitations. First, they show suboptimal performance compared to task-specific models on the training distribution. Second, they suffer from computational inefficiencies as they can only segment one anatomical class per forward pass, requiring multiple passes for multi-class segmentation. Meanwhile they must re-encode context examples for each query image even when using the same

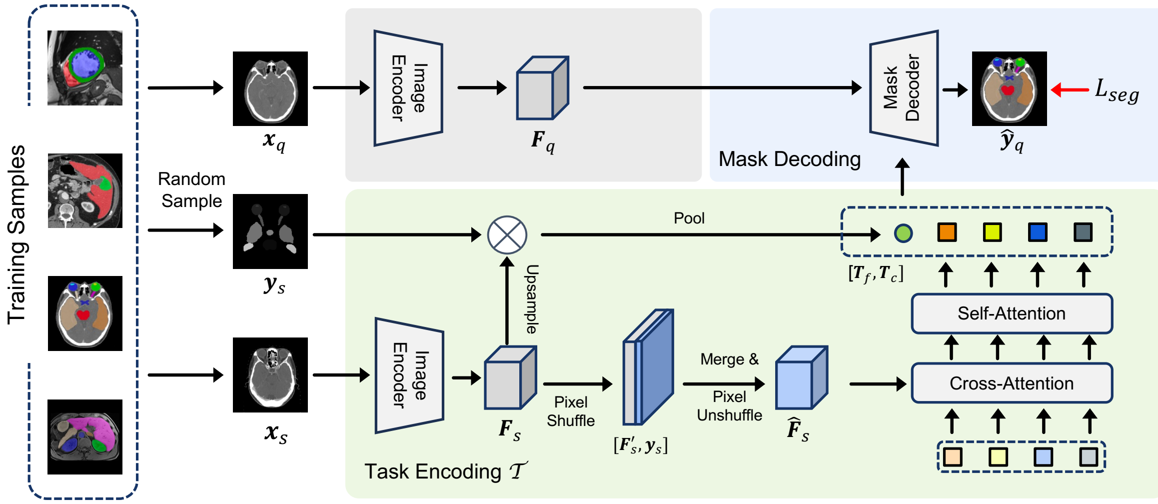

Figure 2. Overview of Iris framework. We design a task encoding module to extract compact task embeddings from reference examples to guide query image segmentation with the mask decoding module, enabling efficient and flexible adaptation to new tasks without finetuning.

reference examples repeatedly. This becomes particularly problematic in high-throughput scenarios where multiple query images need to be processed. In contrast, Iris shows better performance and efficiency. An appealing design of our context task encoding module is to decouple task definition from inference, enabling the encoding of task from reference pairs into task tokens that can be efficiently reused across any number of query images, meanwhile multi-class segmentation can be done within a single forward pass.

The selection of appropriate context examples impacts the performance of in-context learning. Current methods [57] employ image-level retrieval strategies using global image embeddings to find better references. However, this approach faces significant challenges in medical image analysis where each image contains multiple classes including structures like organs, tissues, and lesions. Image-level retrieval inevitably averages features across all structures, leading to a suboptimal reference selection. To address this limitation, Iris introduces an object-level context selection mechanism that enables fine-grained matching of individual classes, focusing on more precise and class-specific reference selection compared to image-level approaches [57].

3. Method

3.1. Problem Definition

Traditional segmentation approaches follow a task-specific paradigm, where each model f_{θ_t} is trained for a specific segmentation task t. Given a dataset $\mathcal{D}_t = \{(\boldsymbol{x}_t^i, \boldsymbol{y}_t^i)\}_{i=1}^{N_t}$ containing N_t image-label pairs, the model learns a direct mapping $f_{\theta_t} : \mathcal{X} \rightarrow \mathcal{Y}$ from the image space $\mathcal{X}$ to the segmentation mask space $\mathcal{Y}$, such that for an image $\boldsymbol{x}_t$, the predicted segmentation mask is given by $\boldsymbol{y}_t = f_{\theta_t}(\boldsymbol{x}_t)$.

In contrast, we formulate a *in-context medical image segmentation* framework. Given a support set $\mathcal{S} = \{(\boldsymbol{x}_s^i, \boldsymbol{y}_s^i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ containing n reference image-label pairs and a query image $\boldsymbol{x}_q \in \mathcal{X}$, a single model f_θ predicts the segmentation mask $\boldsymbol{y}_q$ for the query image conditioned on $\mathcal{S}$:

$$\hat{\boldsymbol{y}}_q = f_\theta(\boldsymbol{x}_q; \mathcal{S}) = f_\theta(\boldsymbol{x}_q; \{(\boldsymbol{x}_s^i, \boldsymbol{y}_s^i)\}_{i=1}^n) \tag{1}$$

For multi-class segmentation tasks, we decompose the problem into multiple binary segmentation tasks.

3.2. Iris Architecture

In Figure 2, Iris introduces a novel in-context learning architecture that decouples task encoding from segmentation inference. This design comprises two key components: (1) a task encoding module that distills task-specific information from reference examples into compact task embeddings, and (2) a mask decoding module that leverages these task embeddings to guide query image segmentation.

3.2.1 Task Encoding Module.

Given a reference 3D image-label pair $(\boldsymbol{x}_s, \boldsymbol{y}_s) \in \mathbb{R}^{D\times H\times W} \times \{0,1\}^{D\times H\times W}$, our task encoding module extracts task representations through two parallel streams to extract comprehensive task representations.

Foreground feature encoding. Medical data volumes present unique challenges in feature extraction due to the presence of fine boundary details and anatomical structures spanning only a tiny portion of voxels. Direct feature pooling at downsampled resolution can lead to information loss or complete disappearance of these critical regions of interest (ROIs). To address this hurdle, we opt in a high-resolution foreground feature encoding process. Given features $\boldsymbol{F}_s \in \mathbb{R}^{C\times d\times h\times w}$ extracted by the encoder E, where

$d = D/r, h = H/r, w = W/r$ are downsampled dimensions with ratio r, we compute the foreground embedding by:

$$\boldsymbol{T}_f = \text{Pool}(\text{Upsample}(\boldsymbol{F}_s) \odot \boldsymbol{y}_s) \in \mathbb{R}^{1\times C} \tag{2}$$

where $\text{Upsample}(\boldsymbol{F}_s) \in \mathbb{R}^{C\times D\times H\times W}$ restores features to the original resolution. By applying the original high-resolution mask $\boldsymbol{y}_s$ directly to the upsampled features, we ensure a precise capture of fine anatomical details and small structures that are vital for medical object segmentation.

Contextual feature encoding. The above encoding process extracted foreground features, but lacks important global context information. We encode these contextual information using learnable query tokens. To efficiently process high-resolution features while managing memory constraints, we employ strategy similar to sub-pixel convolution [42]. For feature map $\boldsymbol{F}_s$, we first expand spatial dimensions while reducing channels:

$$\boldsymbol{F}'_s = \text{PixelShuffle}(\boldsymbol{F}_s) \in \mathbb{R}^{C/r^3\times D\times H\times W} \tag{3}$$

After concatenating with the binary mask $\boldsymbol{y}_s$, we apply a $1\times 1\times 1$ convolution and PixelUnshuffle to return to the original feature resolution:

$$\hat{\boldsymbol{F}}_s = \text{PixelUnshuffle}(\text{Conv}(\text{Concat}[\boldsymbol{F}'_s, \boldsymbol{y}_s])) \in \mathbb{R}^{C\times d\times h\times w} \tag{4}$$

This approach permits a memory-efficient, high-resolution, feature-mask fusion. The merged features $\hat{\boldsymbol{F}}_s$ then interact with m learnable query tokens through cross-attention and self-attention layers to produce contextual embedding $\boldsymbol{T}_c \in \mathbb{R}^{m\times C}$. The final task embedding combines both aspects: $\boldsymbol{T} = [\boldsymbol{T}_f; \boldsymbol{T}_c] \in \mathbb{R}^{(m+1)\times C}$.

For multi-class segmentation, we generate separate task embeddings for each category in $\boldsymbol{y}_s$. This setting maintains a strong efficiency as the computationally intensive feature extraction is shared across classes while the task encoding module remains lightweight.

3.2.2 Mask Decoding Module

The decoder D employs a query-based architecture [8] that efficiently handles both single and multi-class segmentation tasks. For a query image with features $\boldsymbol{F}_q \in \mathbb{R}^{C\times d\times h\times w}$, the task encoding module generates class-specific embeddings $\boldsymbol{T}^k \in \mathbb{R}^{(m+1)\times C}$ for each class k defined in reference image-label pairs. These embeddings are concatenated into a combined task representation $\boldsymbol{T} = [\boldsymbol{T}^1; \boldsymbol{T}^2; ...; \boldsymbol{T}^K] \in \mathbb{R}^{K(m+1)\times C}$, where K is the number of target classes and $K = 1$ for single-class segmentation. The bidirectional cross-attention mechanism processes this representation:

$$\boldsymbol{F}'_q, \boldsymbol{T}' = \text{CrossAttn}(\boldsymbol{F}_q, \boldsymbol{T}) \tag{5}$$

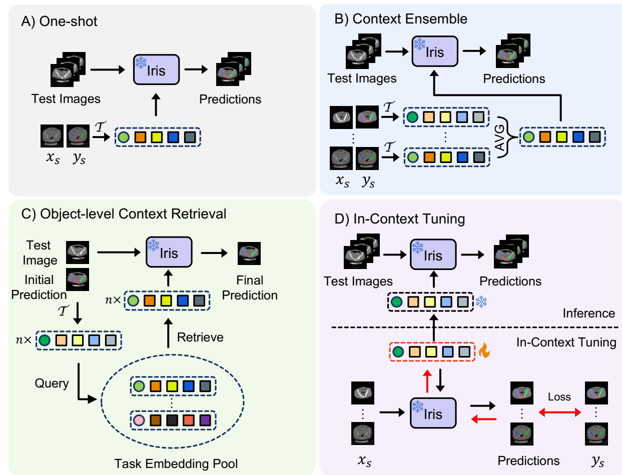

Figure 3. Iris's flexible inference strategies. The red arrows indicates gradient backpropagation.

where $\boldsymbol{F}'_q$ and $\boldsymbol{T}'$ are the updated features. This mechanism enables effective information exchange between class-specific task guidance and query image features. The final segmentation mask is predicted in a single forward pass:

$$\hat{\boldsymbol{y}}_q = D(\boldsymbol{F}'_q, \boldsymbol{T}') \in \{0,1\}^{K\times D\times H\times W} \tag{6}$$

3.2.3 Training.

We train Iris in an end-to-end manner using episodic training to simulate in-context learning scenarios (details in supplementary). Each training episode consists of sampling reference-query pairs from the same dataset, computing task embeddings from the reference pair, and final predicting segmentation for the query image. The model is optimized using a combination of Dice and cross-entropy losses: $\mathcal{L}_{seg} = \mathcal{L}_{dice} + \mathcal{L}_{ce}$. To enhance generalization, we employ data augmentation on both query and reference images, add random perturbation to query images to simulate imperfect references, and randomly drop classes in multi-class datasets to encourage independent class-wise task encoding.

3.3. Flexible Inference Strategies

After training, Iris supports multiple inference strategies suitable for different practical scenarios (see in Figure 3).

Efficient one-shot inference. With just one reference example, Iris first encodes the task into compact embeddings $\boldsymbol{T}$ that can be stored and reused across multiple query images. Unlike major in-context learning methods to recompute contextual information for each query image, our design greatly eliminates redundant computation. Moreover, Iris can segment multiple classes in a single forward pass, contrasting with methods (e.g., UniverSeg [5]) that require separate passes per class. The minimal storage requirement of these embeddings makes Iris particularly desirable for large-scale data processing pipelines.

Table 1. Comparison of segmentation performance across different in-distribution datasets. Values represent mean Dice scores (%).

Method	AMOS CT	AMOS MR	Auto PET	BCV	Brain	CHAOS	KiTS Tumor	LiTS Tumor	MnM	StructSeg H&N	StructSeg Tho	CSI-Wat	AVG
Task-specific Model (Upper Bound)													
nnUNet	88.67	85.42	67.21	83.38	94.12	91.13	81.72	63.11	85.59	78.17	88.53	91.11	83.18
Multi-task Universal Model (Upper Bound)													
Clip-driven	88.95	86.41	70.01	85.03	95.06	91.71	82.73	65.43	86.12	78.44	89.27	90.98	84.18
UniSeg	89.11	86.58	70.09	85.42	95.29	91.83	82.99	65.87	86.29	78.72	89.42	91.23	84.40
Multi-Talent	89.15	86.58	70.89	85.20	95.77	91.38	82.32	65.53	86.30	80.09	89.09	91.32	84.47
Positional Prompt													
SAM	22.23	17.82	20.10	23.34	20.51	20.01	18.21	12.08	10.23	17.23	24.81	13.20	17.97
SAM-Med 2D	50.12	48.66	38.03	50.32	35.28	50.32	30.23	23.27	40.33	39.32	63.87	34.87	40.58
SAM-Med 3D	79.19	76.18	67.14	79.89	42.29	84.79	79.32	32.93	52.67	68.83	83.56	74.23	68.42
In-Context													
SegGPT	45.37	51.78	48.29	49.78	85.27	63.72	40.78	35.98	74.12	40.28	67.28	85.59	57.35
UniverSeg	57.24	52.43	47.23	45.26	87.76	60.46	45.72	36.21	75.24	42.98	66.95	86.68	58.68
Tyche-IS	59.57	54.78	50.98	47.67	89.28	62.73	49.27	37.02	78.92	45.33	69.89	88.99	61.20
Iris (ours)	89.56	86.70	70.02	85.73	96.04	91.85	81.54	65.02	86.08	80.36	89.42	91.97	84.52

Context ensemble. For tasks with multiple reference examples, Iris supports context ensemble for improving performance. We compute task embeddings for each example and average them to create a more robust task representation. This simple averaging strategy combines information from multiple references while maintaining computational efficiency. We extent context ensemble for classes seen during training. Specifically, we maintain a class-specific memory bank that continuously updates task embeddings through exponential moving average (EMA) during the training process. This memory bank stores representative task embeddings for each seen class, enabling direct segmentation for seen classes during inference without requiring context encoding.

Object-level context retrieval. For multi-class segmentation with a pool of reference examples, conventional approaches typically employ image-level retrieval using global embeddings to select semantically similar references [57]. However, this strategy is suboptimal for medical images where multiple anatomical structures coexist, as global embeddings average features across all structures. To enable more precise reference selection, we propose an object-level (class-level) context retrieval strategy. Our approach first encodes class-specific task embeddings for each reference example through our task encoding module - for a reference image with n anatomical classes, we encode n separate task embeddings. For a query image, we obtain initial object segmentation masks using task embeddings from a randomly selected reference. These initial masks are then used to encode n class-specific query task embeddings, which are compared with corresponding reference embeddings in the pool using cosine similarity to select the most similar reference for each class independently. This fine-grained matching allows different structures within the same query image to find their most appropriate references, leading to more accurate segmentation compared to image-level approaches.

In-context tuning. For scenarios requiring adaptation without a full model fine-tuning, Iris offers a lightweight tuning strategy by optimizing only the task embeddings while keeping the model parameters fixed. This tuning process minimizes the segmentation loss between model predictions and ground truth by updating the task embeddings through the gradient descent. In particular, the optimized embeddings can then be stored and reused for similar cases, offering a practical balance between adaptation capability and computational efficiency.

4. Experiment

4.1. Experimental Setup

We evaluate Iris across three key dimensions: in-distribution performance on trained tasks, out-of-distribution generalization to different domains, and adaptability to novel anatomical classes. Additional experiments analyze Iris's computational efficiency, inference strategies, and architectural design choices.

Dataset. Our training data comprises 12 public datasets [3, 6, 15, 18, 21, 22, 25, 26, 35, 40] spanning diverse body regions (head, chest, abdomen), modalities (CT, MRI, PET), and clinical targets (organs, tissues, lesions), split into 75%/5%/20% for train/validation/test. For out-of-distribution evaluation, we use 5 held-out datasets: ACDC [2], SegTHOR [24], and three MRI modalities from IVDM3Seg [16] to evaluate robustness against domain shift; MSD Pancreas (Tumor) [1] and Pelvic1K (Bone) [31] datasets are used for novel class adaptation. We randomly select 20% samples from held-out sets for testing. Detailed dataset information is provided in supplementary materials.

Baseline Models. We compare against four categories of

Table 2. Out-of-distribution comparison on held-out datasets, including generalization capability and performance on unseen classes. Values represent mean Dice scores (%). All in-context models use one-shot inference.

Method	Generalization					Unseen Classes	
	ACDC	SegTHOR	CSI-inn	CSI-opp	CSI-fat	MSD Pancreas	Pelvic
Supervised Upper Bound							
nnUNet	90.97	89.78	91.23	91.04	90.13	54.56	94.73
Task-specific Model							
nnUNet-generalize	82.06	76.92	55.24	85.19	0.23	–	–
Multi-task Universal Model							
CLIP-driven	84.72	78.23	59.73	86.73	1.47	–	–
UniSeg	84.98	78.56	60.02	86.13	1.52	–	–
Multi-Talent	83.79	78.45	58.29	87.01	1.95	–	–
Positional Prompt							
SAM-Med2D	42.23	52.37	29.23	32.71	10.91	10.37	35.71
SAM-Med3D	51.49	68.97	45.32	68.72	23.93	15.83	53.61
In-context							
SegGPT	73.82	60.98	59.87	77.62	35.27	10.67	55.92
UniverSeg	72.43	54.75	63.48	85.32	52.48	10.28	57.81
Tyche-IS	74.91	56.75	64.23	87.13	**55.75**	11.97	61.92
Iris (ours)	**86.45**	**82.77**	**64.44**	**89.13**	47.78	**28.28**	**69.03**

methods: (1) Task-specific models: nnUNet [20]; (2) Universal models: CLIP-driven model [30], UniSeg [50] and Multi-Talent [44]; (3) Foundation models: SAM [23] and its medical variants, SAM-Med2D [9], SAM-Med3D [45]; (4) In-context learning methods: SegGPT [48], UniverSeg [5], and Tyche-IS [39]. All models are trained on our curated dataset, except SAM, with 2D models trained on extracted slices and 3D models with 3D volumes. For SAM-based methods, we simulate user interactions using ground-truth labels during training and evaluation.

Implementation Details. Iris uses a 3D UNet encoder trained from scratch with one-shot learning strategy. We employ the Lamb optimizer [53] with exponential learning rate decay (base lr=2×10^{-3}, weight decay=1×10^{-5}), training for 80K iterations with batch size 32 and 2K warm-up iterations. Data augmentation includes random cropping, affine transformations, and intensity adjustments. Training and inference use $128 \times 128 \times 128$ volume size.

4.2. Comparison with the state-of-the-art

Results on in-distribution classes. We evaluate Iris's performance on twelve diverse medical datasets used during training. As shown in Table 1, Iris achieves state-of-the-art performance with an average Dice score of 84.52%, matching or exceeding task-specific and multi-task models that are optimized for fixed tasks. Existing adaptive approaches show significant limitations. SAM-based methods perform poorly due to their strong reliance on simple positional prompts, with the large performance gap between their 2D and 3D variants (40.58% vs. 68.42%) highlighting the importance of 3D context. Existing in-context learning methods, like SegGPT, UniverSeg, (best: 61.20%) struggle particularly with 3D tasks like AMOS and LiTS due to their 2D architecture, though performing better on 2D-friendly tasks like MnM and CSI-Wat. In contrast, Iris's 3D architecture and efficient task encoding enables consistent high-level performance across all tasks while maintaining its adaptability to unseen novel anatomical classes.

Results on OOD generalization. We evaluate out-of-distribution (OOD) performance on five held-out datasets spanning two types of distribution shifts: cross-center variation (ACDC, SegTHOR) and cross-modality adaptation (CSI variants). In Table 2, Iris demonstrates superior performance across all scenarios, particularly excelling in challenging 3D tasks and large domain shifts.

Both task-specific and multi-task universal models show performance degradation, especially failing catastrophically on CSI-fat with a significant domain gap. While SAM-based methods demonstrate their resilience to domain shifts through strong prior knowledge injected from user interactions, their performance remains limited on the volumetric data. In-context learning methods retain a good performance with cross-modality adaptation (e.g. on CSI-fat), benefiting from the domain-specific knowledge provided by reference examples. However, a 2D-slice-based architecture (e.g., UniverSeg and Tyche) limits its capability on 3D tasks like SegTHOR. In contrast, Iris's task encoding module efficiently extracts and utilizes 3D domain-specific information from the reference examples.

Results on novel classes. To measure the adaptation performance to completely unseen anatomical structures, we evaluate on MSD Pancreas Tumor and Pelvic datasets. Using only one reference example, Iris achieves 28.28% on MSD Pancreas Tumor and 69.03% on Pelvic segmentation, substantially outperforming other adaptive methods (the best competitor: 11.97% and 61.92% respectively). This perfor-

Table 3. Comparison of computational complexity. Empirical measurements of computation on one NVIDIA A100 GPU. We inference with 10 query images and one reference image from AMOS CT dataset with 15 classes. The image size is processed to $128 \times 128 \times 128$ for inference.

Method	Inference Time (s)	Memory (GB)	Parameters (M)
UniverSeg-1	659.4	2.1	1.2
UniverSeg-128	1030.2	12.1	1.2
SAM-Med2D	648.4	1.8	91.1
SAM-Med3D	15.2	2.9	100.5
Iris (Ours)	2.0	7.4	69.4

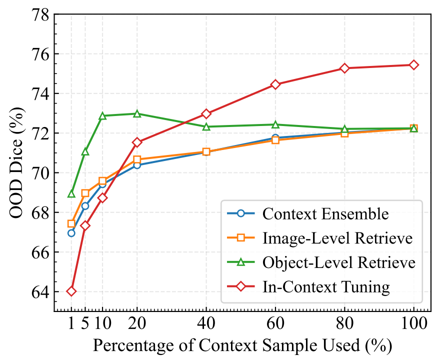

Figure 4. Analysis of different inference strategies.

mance gain is particularly notable given that traditional task-specific models and multi-task models can not handle these novel classes without retraining. These findings demonstrate Iris's strong capability in learning from very limited examples while maintaining meaningful segmentation quality on previously unseen anatomical structures.

Efficiency comparison We analyze computational efficiency for segmenting m query images with n classes using k reference pairs. Iris achieves superior efficiency through two key designs: decoupling task extraction from inference and handling multiple classes in a single forward pass. This results in complexity of $O(k+m)$ compared to $O(kmn)$ in methods like UniverSeg that process each class separately and recompute reference features for every query.

Table 3 compares real-world inference time and memory usage across methods. While UniverSeg's slice-by-slice processing leads to significant overhead with multiple reference slices, and SAM-Med3D requires iterative user interactions (interaction-time not included), Iris efficiently processes entire 3D volumes all at once. For a scenario of segmenting 10 query volumes with 15 classes using one reference volume, Iris completes in 2 seconds. This efficiency advantage grows with more context examples due to Iris's decoupled architecture eliminating redundant reference processing.

4.3. Analysis

Inference strategy. Figure 4 compares our four inference strategies. In this experiment, we maintain a pool of all available context examples and evaluate each strategy's performance as follows.

Context ensemble randomly selects and averages task embeddings from a percentage of the context pool. When using only one context example (1%), it operates as one-shot inference. Performance of context ensemble keeps improving with more context examples and eventually saturates. This strategy is appealing as task embeddings can be precomputed and ensembled into a single robust embedding, enabling inference speed comparable to regular segmentation models.

Both image and object-level retrieval strategies access the entire context pool but utilize only the top-k percent most similar examples as references. While image-level retrieval [57] compares whole-image features and uses all task embeddings from the same retrieved images, object-level retrieval enables more precise reference selection by matching individual classes. Notably, object-level retrieval surpasses full context ensemble performance when using fewer references (e.g., top 10-20%), as it selectively chooses the most relevant examples for each class rather than averaging all available contexts. To validate robustness to initial context selection, we conducted experiments for 10 times with random selection using different percentages of context samples (1%, 5%, 10%), achieving consistent performance (mean and standard deviation: 68.94 ± 0.83, 71.07 ± 0.27, 72.87 ± 0.10 respectively). This strategy is particularly valuable in clinical settings with large patient databases, where retrieving similar cases as references can enhance segmentation accuracy.

In-context tuning optimizes task embeddings initialized from a random reference. While showing a lower performance with limited samples due to overfitting, it achieves positive results with sufficient tuning data. This approach is well suited for scenarios with both a large context pool and available computational resources for fine-tuning.

Overall, Iris offers usable strategies pertinent to different real-world scenarios. Object-level retrieval is designed for high accuracy while requiring access to a large context pool, e.g. a database of previous patient records. Context ensemble offers a strong efficiency of response time. Finally, in-context tuning is applicable when computational resources and sufficient data support are available.

Task embedding analysis. Iris's task encoding module discovers meaningful anatomical relationships without explicit anatomical supervision, learning solely from binary segmentation masks. From Figure 5, the t-SNE visualization of task embeddings reveals natural clustering of anatomical structures that transcends dataset boundaries and imaging modalities. For example, abdominal organs cluster together despite originating from different datasets and modalities (e.g., AMOS-CT, BCV in CT; AMOS-MR, CHAOS in MRI).

We find that feature embeddings capture clinically meaningful anatomical similarities that were never explicitly taught (Figure 5, bottom). Blood vessels like the Inferior

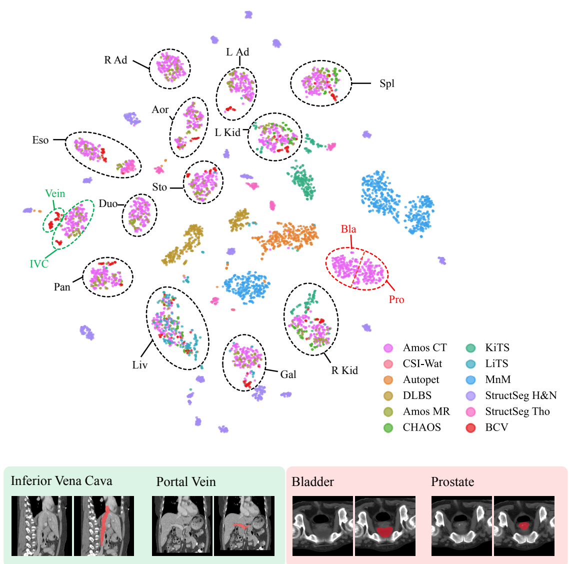

Figure 5. **Top**: Visualizing the task embedding with t-SNE. The color represents dataset, the circle and marks are the classes of the embeddings. **Bottom**: Examples of the similar tasks revealed by the t-SNE result.

Vena Cava (IVC) and Portal/Splenic veins cluster nearby, reflecting their shared tubular structure and similar contrast enhancement patterns in CT. Similarly, the bladder and prostate embeddings show proximity due to their shared soft-tissue characteristics and adjacent anatomical locations. This emergent organization of anatomical concepts demonstrates Iris's ability to automatically distill fundamental anatomical relationships across different segmentation tasks, making it particularly robust for adapting to new anatomical structures.

Generalization performance vs task quantity. We investigate how training data diversity affects Iris's generalization by varying the number of training tasks. From Figure 6 (left), the performance on held-out datasets consistently improves with more training tasks, particularly when the training subset encompasses diverse anatomical structures and imaging modalities. We recognize that models trained on datasets spanning body regions (e.g., brain, chest, and abdomen) show a stronger generalization compared to those trained on narrower anatomical ranges. This finding suggests that an exposure to diverse anatomical patterns is necessary towards more robust and transferable feature learning.

Ablation study. Table 4 analyzes three key components of Iris. High-resolution processing proves crucial for small structures, dramatically improving their segmentation performance from 62.13% to 78.92%. While each component contributes independently, their combination achieves the best results across all metrics, demonstrating substantial improvements over partial implementations. In Figure 6 (right), performance improves with more query tokens but saturates at 10 tokens, which we adopt in our final model to balance performance and efficiency.

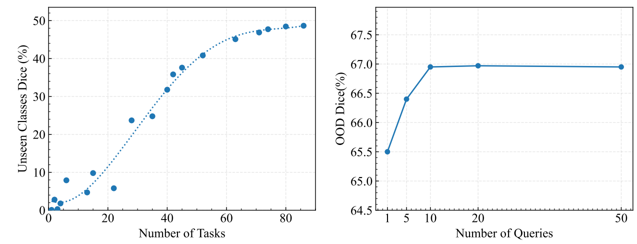

Figure 6. **Left**: Number of tasks used for training v.s. Performance on unseen classes. **Right**: Ablation on the number of queries.

Table 4. Ablation study of different components in Iris. High-Res: high-resolution feature processing; Foreground: foreground feature pooling; Query: query-based contextual encoding.

High-Res	Foreground	Query	In-dist (Avg)	In-dist (Small)	Out-of-dist
	✓	✓	82.10	62.13	62.00
✓	✓		82.47	78.92	65.93
✓		✓	82.06	77.53	64.13
✓	✓	✓	**84.52**	**80.36**	**66.95**

5. Discussion and Conclusion

We introduce Iris as a novel in-context learning framework that enables versatile 3D medical image segmentation through only reference examples. Given just one image-label pair as a reference, Iris can segment arbitrary target classes in test images without any model modification or retraining. Iris reveals strong performance on in-distribution tasks across 12 diverse datasets. Iris's performance is particularly evident to distribution shifts and novel unseen classes on 7 held-out test datasets. The key design of Iris is a decoupled architecture that enables efficient 3D medical image processing and single-pass multi-class segmentation. Iris's inference strategies are suitable for different practical scenarios, from efficient context ensemble-based data processing, high-accuracy object-level context retrieval, to in-context finetuning. Further, Iris's task encoding module offers an appealing means to automatically discover meaningful anatomical relationships purely from segmentation masks, allowing knowledge transfer across different tasks and imaging modalities without explicit anatomical supervision.

Limitations and future work. While Iris demonstrates promising capabilities, several challenges remain to explore. The diversity of training tasks could impact the out-of-distribution generalization, suggesting a critical need for automated methods to create diverse tasks without manual annotation. Although Iris shows strong adaptability to novel tasks, there remains a performance gap with supervised upper bounds in certain scenarios. Future investigation will focus on narrowing this gap and expanding both training and evaluation schemes to cover a broader spectrum of medical imaging applications.

References

[1] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022. 5

[2] 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

[3] 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

[4] 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. 2

[5] Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Universeg: Universal medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21438–21451, 2023. 2, 4, 6

[6] Victor M Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martin-Isla, Alireza Sojoudi, Peter M Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, et al. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. *IEEE Transactions on Medical Imaging*, 40(12):3543–3554, 2021. 5

[7] Qi Chang, Zhennan Yan, Mu Zhou, Di Liu, Khalid Sawalha, Meng Ye, Qilong Zhangli, Mikael Kanski, Subhi Al'Aref, Leon Axel, et al. Deeprecon: Joint 2d cardiac segmentation and 3d volume reconstruction via a structure-specific generative method. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 567–577. Springer, 2022. 1

[8] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit 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. 4

[9] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023. 2, 6

[10] Steffen Czolbe and Adrian V Dalca. Neuralizer: General neuroimage analysis without re-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6217–6230, 2023. 2

[11] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9): 1342–1350, 2018. 1

[12] Konstantin Dmitriev and Arie E Kaufman. Learning multi-class segmentations from single-class datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9501–9511, 2019. 2

[13] Yunhe Gao. Training like a medical resident: Context-prior learning toward universal medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11194–11204, 2024. 1, 2

[14] Yunhe Gao, Mu Zhou, Di Liu, Zhennan Yan, Shaoting Zhang, and Dimitris N Metaxas. A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark. *arXiv preprint arXiv:2203.00131*, 2022. 1

[15] Sergios Gatidis, Tobias Hepp, Marcel Früh, Christian La Fougère, Konstantin Nikolaou, Christina Pfannenberg, Bernhard Schölkopf, Thomas Küstner, Clemens Cyran, and Daniel Rubin. A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. *Scientific Data*, 9(1):601, 2022. 5

[16] Zheng Guoyan, Li Shuo, and Belavy Daniel. Automatic intervertebral disc localization and segmentation from 3d multi-modality mr (m3) images, 2018. 5

[17] Xiaoxiao He, Chaowei Tan, Bo Liu, Liping Si, Weiwu Yao, Liang Zhao, Di Liu, Qilong Zhangli, Qi Chang, Kang Li, et al. Dealing with heterogeneous 3d mr knee images: A federated few-shot learning method with dual knowledge distillation. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023. 1

[18] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019. 5

[19] Rui Huang, Yuanjie Zheng, Zhiqiang Hu, Shaoting Zhang, and Hongsheng Li. Multi-organ segmentation via co-training weight-averaged models from few-organ datasets. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*, pages 146–155. Springer, 2020. 2

[20] 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, 6

[21] 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

[22] A. Emre Kavur, N. Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, Bora Baydar, Dmitry Lachinov, Shuo Han, Josef Pauli, Fabian Isensee, Matthias Perkonigg, Rachana Sathish, Ronnie Rajan, Debdoot

Sheet, Gurbandurdy Dovletov, Oliver Speck, Andreas Nürnberger, Klaus H. Maier-Hein, Gözde Bozdağı Akar, Gözde Ünal, Oğuz Dicle, and M. Alper Selver. CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021. 5

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 6

[24] 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

[25] Bennett Landman, Zhoubing Xu, Juan Lgelsias, Martin Styner, Thomas Langerak, and Klein Arno. Multi-atlas labeling beyond the cranial vault - workshop and challenge, 2020. 5

[26] Hongsheng Li, Jinghao Zhou, Jincheng Deng, and Ming Chen. Automatic structure segmentation for radiotherapy planning challenge 2019, 2019. 5

[27] Di Liu, Jiang Liu, Yihao Liu, Ran Tao, Jerry L Prince, and Aaron Carass. Label super resolution for 3d magnetic resonance images using deformable u-net. In *Medical Imaging 2021: Image Processing*, pages 606–611. SPIE, 2021. 1

[28] Di Liu, Zhennan Yan, Qi Chang, Leon Axel, and Dimitris N Metaxas. Refined deep layer aggregation for multi-disease, multi-view & multi-center cardiac mr segmentation. In *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 315–322. Springer, 2021.

[29] Di Liu, Yunhe Gao, Qilong Zhangli, Ligong Han, Xiaoxiao He, Zhaoyang Xia, Song Wen, Qi Chang, Zhennan Yan, Mu Zhou, et al. Transfusion: multi-view divergent fusion for medical image segmentation with transformers. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 485–495. Springer, 2022. 1

[30] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. *arXiv preprint arXiv:2301.00785*, 2023. 2, 6

[31] Pengbo Liu, Hu Han, Yuanqi Du, Heqin Zhu, Yinhao Li, Feng Gu, Honghu Xiao, Jun Li, Chunpeng Zhao, Li Xiao, et al. Deep learning to segment pelvic bones: large-scale ct datasets and baseline models. *International Journal of Computer Assisted Radiology and Surgery*, 16:749–756, 2021. 5

[32] Xiangbin Liu, Liping Song, Shuai Liu, and Yudong Zhang. A review of deep-learning-based medical image segmentation methods. *Sustainability*, 13(3):1224, 2021. 1

[33] Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. *arXiv preprint arXiv:2305.13310*, 2023. 2

[34] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 2

[35] Carlos Martín-Isla, Víctor M Campello, Cristian Izquierdo, Kaisar Kushibar, Carla Sendra-Balcells, Polyxeni Gkontra, Alireza Sojoudi, Mitchell J Fulton, Tewodros Weldebirhan Arega, Kumaradevan Punithakumar, et al. Deep learning segmentation of the right ventricle in cardiac mri: the m&ms challenge. *IEEE Journal of Biomedical and Health Informatics*, 27(7):3302–3313, 2023. 5

[36] Lingchen Meng, Shiyi Lan, Hengduo Li, Jose M Alvarez, Zuxuan Wu, and Yu-Gang Jiang. Segic: Unleashing the emergent correspondence for in-context segmentation. In *European Conference on Computer Vision*, pages 203–220. Springer, 2024. 2

[37] Ursula Nestle, Stephanie Kremp, Andrea Schaefer-Schuler, Christiane Sebastian-Welsch, Dirk Hellwig, Christian Rübe, and Carl-Martin Kirsch. Comparison of different methods for delineation of 18f-fdg pet–positive tissue for target volume definition in radiotherapy of patients with non–small cell lung cancer. *Journal of nuclear medicine*, 46(8):1342–1348, 2005. 1

[38] Ziwei Niu, Shuyi Ouyang, Shiao Xie, Yen-wei Chen, and Lanfen Lin. A survey on domain generalization for medical image analysis. *arXiv preprint arXiv:2402.05035*, 2024. 1

[39] Marianne Rakic, Hallee E Wong, Jose Javier Gonzalez Ortiz, Beth A Cimini, John V Guttag, and Adrian 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. 2, 6

[40] KM Rodrigue, KM Kennedy, MD Devous, JR Rieck, AC Hebrank, R Diaz-Arrastia, D Mathews, and DC Park. β-amyloid burden in healthy aging: regional distribution and cognitive consequences. *Neurology*, 78(6):387–395, 2012. 5

[41] Wei Shen, Mu Zhou, Feng Yang, Caiyun Yang, and Jie Tian. Multi-scale convolutional neural networks for lung nodule classification. In *Information Processing in Medical Imaging: 24th International Conference, IPMI 2015, Sabhal Mor Ostaig, Isle of Skye, UK, June 28-July 3, 2015, Proceedings 24*, pages 588–599. Springer, 2015. 1

[42] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 4

[43] Yanpeng Sun, Jiahui Chen, Shan Zhang, Xinyu Zhang, Qiang Chen, Gang Zhang, Errui Ding, Jingdong Wang, and Zechao Li. Vrp-sam: Sam with visual reference prompt. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23565–23574, 2024. 2

[44] Constantin Ulrich, Fabian Isensee, Tassilo Wald, Maximilian Zenk, Michael Baumgartner, and Klaus H Maier-Hein. Multitalent: A multi-dataset approach to medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 648–658. Springer, 2023. 2, 6

[45] H Wang et al. Sam-med3d: towards general-purpose segmentation models for volumetric medical images. *Preprint at https://arxiv. org/abs/2310.15161*, 2024. 2, 6

[46] Risheng Wang, Tao Lei, Ruixia Cui, Bingtao Zhang, Hongying Meng, and Asoke K Nandi. Medical image segmentation using deep learning: A survey. *IET image processing*, 16(5): 1243–1267, 2022. 1

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

[48] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023. 2, 6

[49] Yutong Xie, Jianpeng Zhang, Yong Xia, and Chunhua Shen. Learning from partially labeled data for multi-organ and tumor segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[50] Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, and Yong Xia. Uniseg: A prompt-driven universal segmentation model as well as a strong representation learner. *arXiv preprint arXiv:2304.03493*, 2023. 2, 6

[51] Yiwen Ye, Ziyang Chen, Jianpeng Zhang, Yutong Xie, and Yong Xia. Meduniseg: 2d and 3d medical image segmentation via a prompt-driven universal model. *arXiv preprint arXiv:2410.05905*, 2024. 2

[52] Jee Seok Yoon, Kwanseok Oh, Yooseung Shin, Maciej A Mazurowski, and Heung-Il Suk. Domain generalization for medical image analysis: A survey. *arXiv preprint arXiv:2310.08598*, 2023. 1

[53] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019. 6

[54] Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1195–1204, 2021. 2

[55] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Xianzheng Ma, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*, 2023. 2

[56] Yixiao Zhang, Xinyi Li, Huimiao Chen, Alan L Yuille, Yaoyao Liu, and Zongwei Zhou. Continual learning for abdominal multi-organ and tumor segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 35–45. Springer, 2023. 1

[57] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What makes good examples for visual in-context learning? *Advances in Neural Information Processing Systems*, 36:17773–17794, 2023. 3, 5, 7

[58] Yunkun Zhang, Jin Gao, Zheling Tan, Lingfeng Zhou, Kexin Ding, Mu Zhou, Shaoting Zhang, and Dequan Wang. Data-centric foundation models in computational healthcare: A survey. *arXiv preprint arXiv:2401.02458*, 2024. 2

[59] Qilong Zhangli, Jingru Yi, Di Liu, Xiaoxiao He, Zhaoyang Xia, Qi Chang, Ligong Han, Yunhe Gao, Song Wen, Haiming Tang, et al. Region proposal rectification towards robust instance segmentation of biological images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 129–139. Springer, 2022. 1

[60] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2022. 1

[61] Yuyin Zhou, Zhe Li, Song Bai, Chong Wang, Xinlei Chen, Mei Han, Elliot Fishman, and Alan L Yuille. Prior-aware neural network for partially-supervised multi-organ segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10672–10681, 2019. 2