Training Like a Medical Resident: Context-Prior Learning Toward Universal Medical Image Segmentation

Yunhe Gao

Rutgers University

Abstract

*A major focus of clinical imaging workflow is disease diagnosis and management, leading to medical imaging datasets strongly tied to specific clinical objectives. This scenario has led to the prevailing practice of developing task-specific segmentation models, without gaining insights from widespread imaging cohorts. Inspired by the training program of medical radiology residents, we propose a shift towards universal medical image segmentation, a paradigm aiming to build medical image understanding foundation models by leveraging the diversity and commonality across clinical targets, body regions, and imaging modalities. Towards this goal, we develop **Hermes**, a novel context-prior learning approach to address the challenges of data heterogeneity and annotation differences in medical image segmentation. In a large collection of eleven diverse datasets (2,438 3D images) across five modalities (CT, PET, T1, T2 and cine MRI) and multiple body regions, we demonstrate the merit of the universal paradigm over the traditional paradigm on addressing multiple tasks within a single model. By exploiting the synergy across tasks, Hermes achieves state-of-the-art performance on all testing datasets and shows superior model scalability. Results on two additional datasets reveals Hermes' strong performance for transfer learning, incremental learning, and generalization to downstream tasks. Hermes's learned priors demonstrate an appealing trait to reflect the intricate relations among tasks and modalities, which aligns with the established anatomical and imaging principles in radiology. The code is available[1].*

1. Introduction

Medical image segmentation methods generate accurate delineations of anatomical structures which are crucial for disease diagnosis [15, 55] and treatment planning [18, 20, 47]. To date, the prevailing paradigm for medical image segmentation promotes the development of separate models for specific medical objects (e.g., organs or tumors) and image modalities (e.g., CT or MR) [8, 19, 22, 33, 43, 65]. This

---

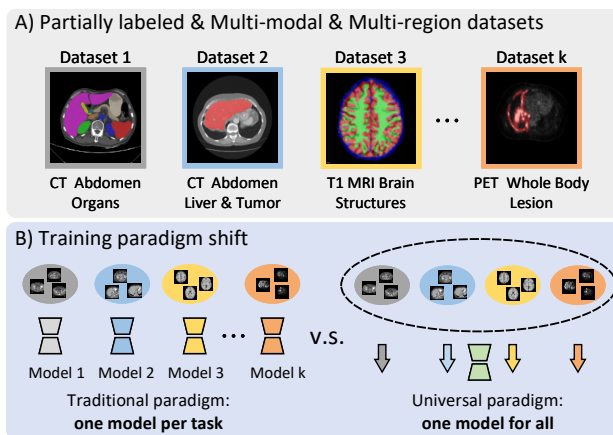[1] https://github.com/yhygao/universal-medical-image-segmentation

Figure 1. A) Clinical diagnostic workflows typically focus on specific specialties and diseases, leading to the curation of image datasets that are partially annotated, multi-modal, and multi-regional. B) Traditional training paradigms involve training separate models for each segmentation task (or dataset). In contrast, we emphasize a universal medical image segmentation paradigm aiming at **one model for all**, leading to a robust and generalizable universal model for diverse tasks.

paradigm is often constrained by the limited training data from the same domain, resulting in compromised model robustness and generalizability. Further scaling up data size for specific segmentation tasks is challenging due to the high cost of data acquisition, collection, and annotation [21, 50]. Moreover, the current paradigm is unable to exploit relationships among medical imaging tasks that are critical for disease understanding. These hurdles together confine the capability and scalability of medical image segmentation models.

Inspired by radiology residency programs [1, 2, 53], we recognize that radiologists' expertise arises from routine exposure to wide-ranging images across body regions, diseases, and imaging modalities. Despite the fact that the human body exhibits anatomical variability, it is fundamentally composed of various types of tissues and structures whose appearance in images are often statistically similar [54]. These tissues display specific visual characteristics under different imaging modalities [4, 31]. During the training process,

residents acquire prior knowledge about the regions of interest (ROI) across imaging modalities for interpreting various types of patterns. Inspired by this observation, we prioritize a paradigm shift toward universal medical image segmentation, which seeks to harness the diversity and commonality of medical images, to build a comprehensive, unified segmentation model. However, addressing the heterogeneity among medical imaging tasks faces daunting challenges.

First, each image dataset has partial and incomplete ROI annotations due to their distinct needs and clinical objectives. As seen in Figure 1, dataset 1 contains annotations for 15 abdominal organs, while dataset 2 is only annotated for the liver and tumor. Secondly, class definitions can vary depending on the clinical target. To illustrate this point, in dataset 1, the liver is annotated as a whole organ, whereas in dataset 2, the same liver region is divided into two categories of liver and tumor. Additionally, images from different datasets can exhibit significant statistical divergence due to factors such as imaging modalities and body regions (dataset 3 and k).

Given these challenges, intriguing questions remain to be fully explored. How can segmentation tasks with different target ROIs mutually benefit from each other? Is the underlying feature representation transferable across different body regions? Despite the visual difference in imaging modalities, how can a model discern and utilize the meaningful commonalities between them?

In this paper, we strive to answer the above questions. Our main contributions are:

- By exploring the universal medical image segmentation paradigm, we show a single unified model can handle tasks across various ROIs, anatomical regions, and modalities.
- We introduce a novel context-prior pool to learn two important types of prior knowledge, *task* and *modality*, directly from medical images. Different from using one-hot [63] or CLIP embeddings [44] to inject task information, Hermes's learned priors are able to capture complex relations among tasks and modalities that are aligned with established anatomical and imaging principles.
- Through a carefully curated eleven datasets, our systematic analysis reveals the strong capability of Hermes in accuracy and model scalability.
- With two additional downstream datasets, we show Hermes's superior performance in transfer learning, incremental learning, and generalization, affirming the efficacy of the universal paradigm in acquiring robust and generalizable image representations.

2. Related Work

Partially labeled data problem. Disease-specific clinical workflows often yield datasets with single-modality images and partial annotations tailored to specific clinical objectives. Early efforts aimed to combine multiple datasets by conditioning on labels [16], regularizing organ size distribution [66], or generating multi-organ pseudo labels for co-training [30]. Recent approaches propose different conditioning methods to inject task information. DoDNet [63] embeds the task index with one-hot vectors as additional model inputs, but these one-hot task vectors are orthogonal with no task relations embedded. CLIP-driven universal model [44] conditions on CLIP text embeddings, but the CLIP encoder, rarely trained with medical data, struggles with the semantics and relationships of complex medical concepts (Fig. 4, 5). MultiTalent [58] uses multiple task-specific heads. Moreover, DoDNet, CLIP-driven and MultiTalent limit their scope to a single body region or imaging modality. UniSeg [61] introduces novel learnable task prompts but does not consider the modality information. In contrast, our work focuses on learning priors directly from a diverse array of medical data sources by addressing both task and modality heterogeneity simultaneously.

Prior knowledge in medical image analysis. Prior knowledge typically involves the understanding of anatomical structures or specifics of imaging modalities before further processing. Traditionally, atlas-based techniques [6, 32] and statistical shape models [14], leverage prior anatomical knowledge by aligning a predefined atlas or template to a patient's image. In deep learning, major methods embed prior knowledge into network architectures or training strategies. For example, integrating graphical models to embed spatial and anatomical knowledge [11, 35], or introducing a shape prior knowledge regularization loss through an autoencoder [20, 48]. Our work diverges by harnessing context-prior learning to learn task and modality knowledge directly from medical data, where this prior knowledge is injected into the backbone to enhance segmentation during inference.

Universal image segmentation aims to unify semantic, instance, and panoptic segmentation into one framework [12, 13, 64]. However, in 3D medical imaging, it is uncommon to find multiple instances of the same object. In addition, fully annotated medical datasets are rarely available in the community, making instance and panoptic segmentation highly unsuitable. As a result, universal medical image segmentation focuses on the semantic segmentation of medical objects. While sharing similarities, developing universal medical image segmentation possesses unique challenges, including the presence of partial annotation, conflicting class definitions, and heterogeneous medical-image contents [50].

3. Method

3.1. Preliminary

Problem definition. We hereby define *universal medical image segmentation* as the endeavor to train a universal model that learns from diverse medical datasets and performs various image segmentation tasks. Given a set of K datasets, let
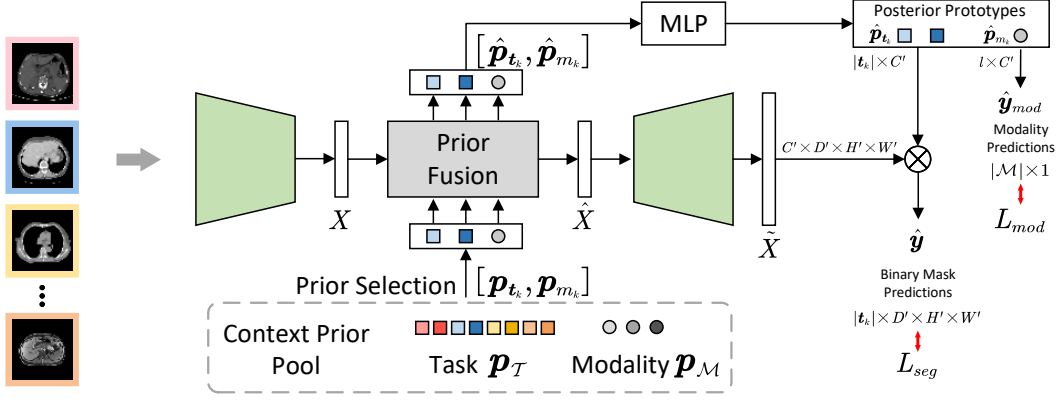
Figure 2. Illustration of Hermes. A context-prior knowledge pool, including task and modality priors, is learned with the backbone. Through oracle-guided selection and combination of these priors, Hermes can address a variety of segmentation tasks and image modalities.

$\mathcal{D}_k = \{(\boldsymbol{x}_k^i, \boldsymbol{y}_k^i, m_k, \boldsymbol{t}_k)\}_{i=1}^{N_k}$ be the k-th dataset comprised of N_k data pairs, where each data pair $(\boldsymbol{x}_k^i, \boldsymbol{y}_k^i) \in (\mathcal{X}_k \times \mathcal{Y}_k)$. m_k denotes the imaging modality of $\mathcal{D}_k$ and $\boldsymbol{t}_k$ represents the tasks contained in $\mathcal{D}_k$, where each task represents the segmentation of a unique clinical target annotated in $\mathcal{Y}_k$, and $|\boldsymbol{t}_k| = |\mathcal{Y}_k|$. The collective dataset is given by $\mathcal{D} = \cup_{k=1}^K \mathcal{D}_k$, the corresponding collective tasks are $\mathcal{T} = \cup_{k=1}^K \boldsymbol{t}_k$, and the collective modalities are $\mathcal{M} = \cup_{k=1}^K m_k$. The objective is to train a single model $f_\theta : \mathcal{X} \to \mathcal{Y}$, parameterized by θ, where $\mathcal{X} = \cup_{k=1}^K \mathcal{X}_k$ and $\mathcal{Y} = \cup_{k=1}^K \mathcal{Y}_k$.

3.2. Oracle-guided context-prior learning

In this section, we introduce Hermes, an oracle-guided context-prior learning approach, as depicted in Fig. 2. Drawing inspiration from radiology residency training, Hermes explicitly learns context-prior knowledge along with the segmentation backbone from diverse medical imaging sources. In line with radiologists' practice for image interpretation, Hermes uses the tasks to be processed $\boldsymbol{t}$, and the imaging modality m as guiding oracles during inference. These oracles guide the selection and combination of two types of priors - *task* and *modality* - to aid the backbone in addressing any specified tasks from the collective dataset. As a model-agnostic approach, Hermes can be seamlessly integrated with existing segmentation backbones, offering a versatile solution for universal medical image segmentation.

Task context prior. To learn from varied datasets and handle the incomplete annotation and potential conflict of class definition, each task in $\mathcal{T}$ is formed as a binary segmentation task. For example, if ROIs in different datasets share the exact same definition, they are considered a single task, otherwise are treated as separate tasks. Each task is associated with a unique task prior, which is implemented as a learnable token. For the collective dataset $\mathcal{D}$ comprising $|\mathcal{T}|$ tasks, we initialize a task prior pool $\boldsymbol{p}_\mathcal{T} \in \mathbb{R}^{|\mathcal{T}| \times C}$, where C denotes the token dimension. Given a training image from $\mathcal{D}_k$, we

use the task IDs $\boldsymbol{t}_k$ as an oracle to guide the selection of corresponding task priors $\boldsymbol{p}_{\boldsymbol{t}_k} \in \mathbb{R}^{|\boldsymbol{t}_k| \times C}$ from the pool, on which the model is conditioned to complete specified tasks. This design enables Hermes to flexibly select and combine task priors based on clinical objectives, accommodating a wide array of medical segmentation tasks.

Modality context prior. Medical images come with a variety of imaging modalities, each possessing distinct image attributes, intensity profiles, and noise patterns. To reduce the modeling difficulty, we introduce a modality prior for each modality. Similar to the task prior, we initialize a modality prior pool $\boldsymbol{p}_\mathcal{M} \in \mathbb{R}^{|\mathcal{M}| \times l \times C}$, where l denotes the length of the token, and C is the dimension of each token. We employ multiple tokens of length l for each modality. When an image with modality m_k is processed, we select the corresponding modality prior token $\boldsymbol{p}_{m_k} \in \mathbb{R}^{l \times C}$ and concatenate it with the task prior tokens: $\boldsymbol{p} = [\boldsymbol{p}_{\boldsymbol{t}_k}, \boldsymbol{p}_{m_k}] \in \mathbb{R}^{(|\boldsymbol{t}_k|+l) \times C}$.

Conditioned segmentation. Given an encoded image feature map $X \in \mathbb{R}^{C \times D \times H \times W}$ from the segmentation backbone, where D, H, W represent the 3D feature map sizes, a prior fusion module is applied to adaptively fuse context-prior tokens and image features:

$$\hat{\boldsymbol{p}}, \hat{X} = Fusion(\boldsymbol{p}, X). \qquad (1)$$

Here, $\hat{\boldsymbol{p}} = [\hat{\boldsymbol{p}}_{\boldsymbol{t}_k}, \hat{\boldsymbol{p}}_{m_k}]$ represents the posterior tokens and and $\hat{X}$ is the prior-injected feature map. The prior fusion module can be implemented with the attention mechanism. For CNN backbones, we employ bi-directional cross-attention modules. For Transformer, we merge the prior tokens into the attention module (detailed in the supplementary). The posterior tokens, processed through a multilayer perceptron (MLP), yield posterior prototypes serving as class classifiers. The predicted probability distribution for each task is computed via the inner product of these posterior task prototypes and image features, followed by a sigmoid function:

$$\hat{\boldsymbol{y}} = p_\theta(y_k = 1 | \boldsymbol{x}, \boldsymbol{p}_{\boldsymbol{t}_k}) = \sigma(\langle \hat{\boldsymbol{p}}_{\boldsymbol{t}_k}, \tilde{X} \rangle), \qquad (2)$$

where $\hat{\boldsymbol{p}}_{t_k} \in \mathbb{R}^{|\boldsymbol{t}_k| \times C'}$ is the posterior prototypes (overload the $\hat{\boldsymbol{p}}_{t_k}$ in equation (1)), $\tilde{X} \in \mathbb{R}^{C' \times D' \times H' \times W'}$ is the output feature map of the decoder, $\hat{\boldsymbol{y}} \in \mathbb{R}^{|\boldsymbol{t}_k| \times D' \times H' \times W'}$ is the predicted binary probability map for tasks $\boldsymbol{t}_k$. To help modality knowledge learning, we use auxiliary supervision on modality priors. Specifically, we predict image modality using posterior modality prototypes via global average pooling and a linear layer: $\hat{\boldsymbol{y}}_{mod} = \text{Linear}(\text{GAP}(\hat{\boldsymbol{p}}_{m_k}))$.

During training, Hermes learns class-specific representations for each task prior token and typical image features for each modality prior. In inference, context-priors and image feature maps undergo a bi-directional update, where priors assimilate instance-specific information from feature maps, and in turn, feature maps are augmented with learned prior knowledge. This interactive process is realized by attention weights, making Hermes a dynamic and adaptive model that enhances segmentation performance.

Hierarchical modeling. The effectiveness of hierarchical modeling is well-established in dense prediction tasks [9, 42]. We apply our context-prior approach hierarchically at multiple scales. The posterior prototypes from each scale are concatenated together and processed by an MLP as the final prototypes for segmentation. This design allows Hermes to learn prior knowledge across different scales, effectively merging multi-scale contextual information to improve the segmentation performance.

Algorithm 1 Hermes Training

1: **Input:** Training dataset $\mathcal{D} = \cup_{k=1}^{K} \mathcal{D}_k$, where $\mathcal{D}_k = \{(\boldsymbol{x}_k^i, \boldsymbol{y}_k^i, m_k^i, \boldsymbol{t}_k^i)\}_{i=1}^{N_k}$. Randomly initialized segmentation backbone f_θ, task and modality prior pool $\boldsymbol{p}_\mathcal{T}$ and $\boldsymbol{p}_\mathcal{M}$
2: **while** *not converged* **do**
3:     **for** j in $[1, 2, \ldots,$ batch size$]$ **do**
4:         Randomly choose a training sample $\boldsymbol{x}, \boldsymbol{y}$ from $\mathcal{D}$
5:         Task and modality priors selection: $\boldsymbol{p}_t$ and $\boldsymbol{p}_m$
6:         Prior concatenation: $\boldsymbol{p} = [\boldsymbol{p}_t, \boldsymbol{p}_m]$
7:     **end for**
8:     Assemble training mini-batch $\mathcal{B}$
9:     Predict segmentation and modality: $\hat{\boldsymbol{y}}, \hat{\boldsymbol{y}}_{mod} = f_\theta(\boldsymbol{x}, \boldsymbol{p})$
10:    Compute loss: $L_{seg} + \lambda L_{mod}$
11:    Update f_θ, $\boldsymbol{p}_\mathcal{T}$ and $\boldsymbol{p}_\mathcal{M}$
12: **end while**

Training and losses. Hermes employs a joint training strategy, using mixed batch training that incorporates samples from multiple datasets within each mini-batch (see Algorithm 1). The primary segmentation loss is a combination of binary cross-entropy and Dice loss: $L_{seg} = L_{bce}(\boldsymbol{y}, \hat{\boldsymbol{y}}) + L_{dice}(\boldsymbol{y}, \hat{\boldsymbol{y}})$, The auxiliary modality loss uses cross-entropy loss on the predicted modality $\hat{\boldsymbol{y}}_{mod}$ and ground truth modality $\boldsymbol{m}$: $L_{mod} = L_{ce}(\hat{\boldsymbol{y}}_{mod}, m)$. The backbone and priors are jointly optimized to synergize representation learning and context-prior knowledge learning.

4. Results

4.1. Experiments setup

Dataset selection and experiment design. Our collective training dataset, detailed in Table 1, comprises eleven public datasets, chosen for their variety in body regions, modalities, and clinical targets. We assess model scalability by exploring the effect of model size on performance in both traditional and universal segmentation approaches. In addition, we further evaluate Hermes' capability in downstream tasks under the settings of transfer learning, incremental learning and generalization with two additional datasets. Finally, we show the priors learned by Hermes are able to accurately capture the complex relationships between tasks and modalities.

Standardized preprocessing. Training with diverse and heterogeneous data is a non-trivial problem. We implement a standardized preprocessing pipeline for all datasets. Firstly, we align all images to the same coordinate system and re-sample the spacing to a uniform $1.5 \times 1.5 \times 1.5 \, mm$. Subsequently, we normalize image intensities; for CT data, we employ a clipping window of [-990, 500], while for MR and PET data, we clip at the 2nd and 98th percentiles of the intensity distribution. Finally, we conduct z-score normalization on each image, ensuring that all data exhibits a zero mean and unit standard deviation.

Implementation details. We implement Hermes using Py-Torch and train the model with a batch size of 16 over 200 epochs. We use the LAMB [62] optimizer with a learning rate of 0.002 and an exponential learning rate decay. Data augmentations, including random cropping, rotation, scaling, brightness, contrast, and gamma perturbation, are applied on the fly during training. We use a patch size of $128 \times 128 \times 128$ for 3D training. The context-prior learning method is applied at the scales of $4\times, 8\times$, and $16\times$ down-sampling. Given the variation in annotation styles across datasets, Classes with identical names across datasets are considered unique ROIs, except in AMOS CT and AMOS MR, resulting in 71 task prior tokens. Different MRI sequences are treated as distinct modalities, leading to a total of five modalities (CT, T1 MRI, T2 MRI, cineMRI, and PET) in the modality prior pool, with the length l set to 10. We set $\lambda = 0.001$ for the auxiliary modality loss as it is much easier than segmentation. The details of the datasets and train/val/test split are available in the supplementary.

4.2. Results

We start initial experiments under the traditional training paradigm by training individual models for each dataset. We use several representative medical image segmentation backbones, including CNN models nnUNet [33], UNet [52] with residual building blocks [26]: ResUNet, and Transformer models SwinUNETR [25] and MedFormer [23]. From Table 2, ResUNet and MedFormer demonstrate slightly better

Table 1. Datasets statistics. The upper datasets are for upstream training and analysis. The bottom two datasets are for downstream tasks on transfer learning, incremental learning, and generalization.

Dataset	Body Region	Modality	Clinical Target	#Cls	Size
BCV [40]	Abdomen	CT	Organs	13	30
LiTS [5]	Abdomen	CT	Liver & Tumor	2	131
KiTS [29]	Abdomen	CT	Kidney & Tumor	2	210
AMOS CT [34]	Abdomen	CT	Organs	15	300
SS T [41]	Thorax	CT	Organs	6	50
SS H [41]	Head & Neck	CT	Organs	22	50
AMOS MR [34]	Abdomen	MRI	Organs	13	60
CHAOS [37]	Abdomen	T1 & T2 MRI	Organs	4	60
M&Ms [7]	Cardiac	cineMRI	Structures	3	320
DLBS [51]	Brain	T1 MRI	Structures	3	213
AutoPET [24]	Whole body	PET	Lesions	1	1014
SegTHOR [39]	Thorax	CT	Organs	3	40
MSD Pancreas [3]	Abdomen	CT	Pancreas & Tumor	2	281

results under the traditional paradigm. Therefore, in subsequent experiments under the universal paradigm, we employ ResUNet and MedFormer as representative backbones for CNN and Transformer to implement Hermes, i.e. Hermes-R and Hermes-M. We also compare against large-scale self-supervised pretraining and finetuning (SSL) methods: DeSD [60] (pretrained on 10,594 CT scans) and UniMiss [59] (pretrained with 5,022 CT scans and 108,948 2D images).

Results on target ROIs, body regions, and image modalities. From Table 2, we see that: **First**, a mutual enhancement in performance is observed across different target ROIs. Specifically, Hermes-R shows significant gains over ResUNet in tumor and lesion segmentation in the LiTS and KiTS datasets, increasing by 2.62% and 3.57%, respectively. Across the other ROIs in eleven datasets, Hermes consistently improves results, with Hermes-R outperforming ResUNet by 2.28%, and Hermes-M surpassing MedFormer by 1.89%. **Second**, there's a clear transferability of underlying representations between different body regions. For example, in thoracic and head&neck regions represented by SS T and SS H datasets, Hermes-R enhances segmentation in thoracic targets by 0.91% in SS T and by 2.5% in SS H compared to ResUNet. **Third**, the commonality between imaging modalities can be harnessed to enhance performance, even though they possess entirely different visual characteristics. For instance, despite the inherent image differences between PET to other MRI and CT datasets, Hermes-R has a remarkable improvement of 8.17% with the universal paradigm. **Lastly**, although DeSD and UniMiss are pretrained on much larger datasets than Hermes, they do not transfer well to different tasks, especially cross-modalities. For example, DeSD is notably worse in the DLBS with brain T1 MRI images, and UniMiss underperforms in AutoPET.

Comparison with other methods under the universal paradigm. We adapt and extend several previous methods to fit the universal paradigm. Multi-decoder [10] uses a shared encoder and separate decoders per dataset. DoDNet [63] was

originally designed for organ and tumor segmentation in CT images with one-hot task vector embeddings; we broaden it into multi-modality segmentation. CLIP-Driven [44] was intended for partially labeled CT image analysis using CLIP text encoder; we adapt it by including modality information in its text prompt: "*A CT/MRI/PET of a [CLS].*" MultiTalent [58] uses multiple segmentation heads. UniSeg [61] uses learnable task prompts. All methods were implemented using the same ResUNet backbone for fairness.

Figure 3 (A) presents a comprehensive comparison of traditional and universal paradigms from six perspectives, detailed numbers are in supplementary. All universal-trained methods demonstrate consistent improvement over the traditional paradigm, particularly for difficult classes, underscoring the universal paradigm's potency in deriving robust representations. Within the universal paradigm, Hermes outperforms other methods in all six perspectives. For example, Multi-Decoder's scalability is limited by the linear growth of decoders with dataset quantity, and it only allows knowledge sharing at the encoder level. DoDNet's one-hot task embedding is incapable of capturing inter-task relationships. CLIP-Driven, despite incorporating knowledge from text prompts, can't effectively encode discriminative information for medical tasks, as its CLIP text encoder lacks training on medical texts, leading to subpar performance, especially for tumors&lesions. MultiTalent demonstrates good performance with task-specific segmentation heads, yet it does not reach the effectiveness of Hermes because it lacks the integration of prior knowledge during inference. Although UniSeg incorporates learnable task prompt maps, it overlooks modality information and learns prompts at the lowest resolution, missing critical multi-scale information. In contrast, Hermes possesses two key advances: (1) It utilizes versatile task and modality context priors, learned directly from medical data, which accurately capture the nuances of task and modality knowledge, as demonstrated in Fig. 4 and 5. (2) It employs hierarchical modeling to help multi-scale learning of prior knowledge.

Model scalability. In Figure 3 (B), we find that, in contrast to neural scaling law [17, 26, 36, 45], increasing the model scale yields marginal performance gain in the traditional paradigm due to limited size and diversity of individual dataset. The limited data from the same distribution cannot fully utilize the larger model's capacity, potentially leading to overfitting. On the contrary, the universal paradigm shifts this dynamic, with an increased backbone scale for Hermes leading to a remarkable performance gain. This finding suggests the proposed universal paradigm's ability to harness the potential of larger models with diverse data and tasks. This approach paves the way for training large-scale models in medical imaging, a feat that is challenging under traditional training paradigms.

Table 2. Universal segmentation results measured with Dice score (%). The upper table presents models with CNN backbones. The lower table shows models with Transformer backbones. Hermes-R denotes our method using ResUNet backbone, while Hermes-M uses MedFormer backbone.

| Paradigm	Model	BCV	SS T	SS H	LiTS T	KiTS T	AMOS CT	AMOS MR	CHAOS	M&Ms	AutoPET	DLBS	AVG
Traditional	nnUNet [33]	84.23	88.53	78.17	64.91	81.72	88.79	85.49	91.34	85.65	65.43	94.22	82.59
Traditional	ResUNet	84.36	88.59	78.12	64.87	81.89	88.97	85.43	91.41	85.74	65.52	94.31	82.65
SSL	DeSD [60]	83.62	88.11	76.56	64.43	82.52	86.36	82.56	91.55	86.46	69.02	86.81	81.64
Universal	Hermes-R	**85.98**	**89.50**	**80.62**	**67.49**	**85.46**	**89.63**	**86.78**	**92.01**	**86.94**	**73.69**	**96.21**	**84.93**
Traditional	SwinUNETR [25]	83.32	88.36	72.21	64.82	74.32	88.29	83.97	88.34	83.28	64.39	92.01	80.32
Traditional	MedFormer [23]	84.61	89.04	78.71	66.24	82.09	89.45	85.58	91.85	86.02	66.01	95.13	83.39
SSL	UniMiSS [59]	84.97	88.29	77.41	63.94	61.21	85.82	83.51	91.35	85.75	60.32	95.35	80.64
Universal	Hermes-M	**86.29**	**89.61**	**81.19**	**68.32**	**85.98**	**89.98**	**87.20**	**92.22**	**87.02**	**74.04**	**96.54**	**85.28**
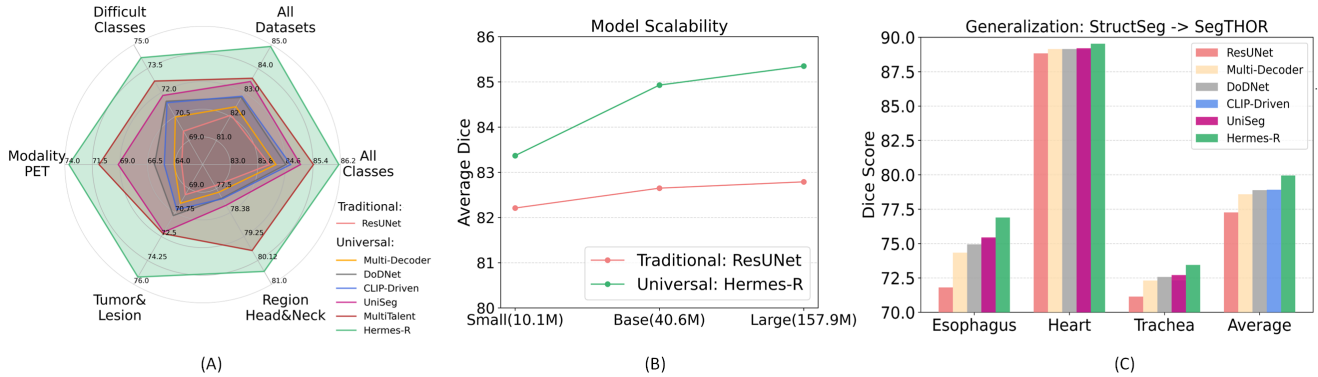
Figure 3. (A) Comparison with other SOTA methods. ROIs with Dice scores lower than 80 under the traditional paradigm are defined as 'difficult classes'. (B) Model scalability analysis. We scale ResUNet down and up to three variants: ResUNet-Small (10.1M), ResUNet-Base (40.6M), and ResUNet-Large (157.9M), and the same for Hermes. All other experiments use ResUNet-Base as the backbone unless specified. (C) Generalization from StructSeg to SegTHOR.

4.3. Downstream tasks

In the universal paradigm, the upstream multi-task training encourages the models to learn robust and generalizable representations across various tasks. We extend to perform an analysis of downstream tasks in the following section.

Transfer learning. We perform evaluation on the challenging MSD pancreas & tumor dataset [3], divided into training (214 samples), validation (10 samples), and testing (57 samples) sets. To assess the influence of downstream data volume, models are fine-tuned on 1%, 10%, 50%, and 100% of the training samples. We report average Dice over 5 runs to reduce variability from the training samples.

Table 3 compares the efficacy of different transfering methods. Traditional single-dataset pretraining with AMOS CT (15 organs) or KiTS (kidney & tumor) shows improved performance with a small amount of downstream data (1%-10%), but these gains saturate as the volume of downstream task data increases (50%-100%). Notably, KiTS pretraining even lags behind training from scratch when using the full data. In contrast, the improvements offered by the universal paradigm, including DoDNet, CLIP-Driven, and Hermes, consistently outperform single-dataset transfer. Despite DeSD's self-supervised pretraining on a larger dataset, Hermes exhibits superior transferability across all data volumes.

Hermes particularly excels in both constrained (1%-10%) and abundant (50%-100%) data scenarios, demonstrating the value of upstream task diversity in enhancing transfer learning for downstream medical image segmentation tasks.

Incremental learning. We test model performance in a more challenging scenario that requires the model to retain knowledge from previous tasks while learning new ones. For all four compared methods, incremental learning is accomplished by keeping the backbone fixed and finetuning only the new task conditions. Table 3 shows that Hermes excels in this setting, particularly under limited data conditions (1%-10%), even outperforming transfer learning results from AMOS CT and KiTS. With ample data, Hermes maintains competitive performance comparable to that of full model tuning. This capability highlights the strength of the backbone's representation learning, fostered during diverse upstream universal training, and shows Hermes' adaptability to new downstream medical tasks.

Generalization. We directly apply models trained on the StructSeg to the SegTHOR datasets, both of which contain thoracic CT scans, and evaluate the performance on three overlapping categories as seen in Figure 3 (C). Compared to ResUNet based on the single-dataset training, all methods trained under the universal paradigm show better generalization, with Hermes leading the performance. This outcome

Table 3. Transfer and incremental learning on the MSD Pancreas & Tumor [3] dataset

| Setting	Model	1%		10%		50%		100%	
		Pan	Tumor	Pan	Tumor	Pan	Tumor	Pan	Tumor
Scratch	ResUNet	44.60	7.67	74.47	23.90	78.89	44.52	80.45	51.06
Transfer	ResUNet (AMOS CT)	56.08	8.31	77.15	25.53	80.53	46.16	81.23	52.21
	ResUNet (KiTS)	52.68	9.28	75.11	27.33	79.07	45.72	79.23	50.64
	DeSD [60] (10,594 CT)	67.82	13.89	78.11	35.82	80.95	50.23	81.97	59.11
	DoDNet [63]	66.62	11.97	76.83	31.92	80.82	47.79	81.41	53.62
	CLIP-Driven [44]	67.95	12.12	77.49	32.37	80.92	48.92	81.45	54.71
	UniSeg [61]	69.05	12.35	77.33	33.87	80.93	49.63	81.96	55.58
	Hermes-R	**72.71**	**16.73**	**79.12**	**44.31**	**81.14**	**55.31**	**82.73**	**61.41**
Incremental	DoDNet [63]	64.36	7.91	65.78	14.39	69.28	21.97	69.87	22.32
	CLIP-Driven [44]	71.52	10.79	75.41	23.22	76.25	29.02	74.72	30.99
	UniSeg [61]	71.68	12.31	72.96	19.82	76.95	27.02	74.40	28.22
	Hermes-R	**72.69**	**15.52**	**76.89**	**28.61**	**79.44**	**43.12**	**79.98**	**47.12**
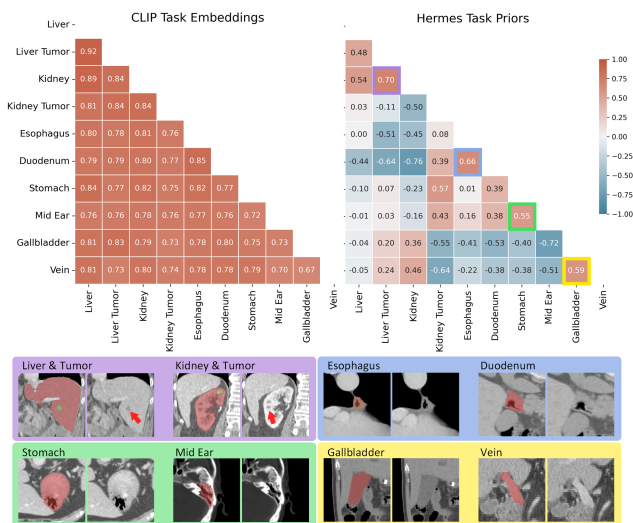
Figure 4. Upper: Cosine similarity of Hermes's learned task priors and CLIP's task embeddings. Lower: Example structures that have high similarity. Hermes's priors are learned directly from medical data and are able to capture intricate relationships among tasks, while CLIP's embeddings tend to encode all objects into similar embeddings, resulting in a loss of discriminative details.

suggests that even with StructSeg being the only thoracic data in our upstream universal training, the inclusion of diverse data enhances potential generalization capabilities.

4.4. Analysis

Learned prior analysis. We illustrate how Hermes adeptly learns task and modality context priors that are consistent with the established anatomical and imaging principles. In Fig. 4, we show the cosine similarity between task priors learned with Hermes and the one of CLIP text embeddings used in [44]. Note that this is pure prior knowledge without considering any image features. The CLIP text encoder, rarely trained with medical data, tends to encode all medical objects with high cosine similarity, offering limited prior knowledge about tasks. In contrast, **Hermes excels in reflecting the intricate relationships among tasks by**

Table 4. Ablation study

Design	Avg Dice
Task prior only single scale	81.72
Task prior only + hierarchy	82.98
Task prior only - unidirectional fusion	82.04
Task + modality priors	83.21
Task + modality priors + modality auxiliary loss	83.37

learning directly from medical images. This is evident in how objects with similar visual features show high similarity, and vice versa. For example, liver tumors display a notable similarity with kidneys, attributable to their shared visual features with the renal medulla or pelvis. Also, the duodenum and esophagus, being parts of the digestive system, demonstrate similarity in their tubular structures filled with air. Intriguingly, despite their significant anatomical and functional differences, the middle ear and stomach are shown to have comparable densities in CT scans, attributed to their air-filled hollows. In contrast, fluid-filled structures like the gallbladder and veins demonstrate higher similarity to each other but differ markedly from hollow organs such as the esophagus, stomach, or middle ear. **Hermes also captures accurate prior knowledge about imaging modalities**. As illustrated in Fig. 5, Hermes effectively discerns distinct characteristics of various imaging modalities. Hermes identifies CT as quite distinct from other modalities like MRI and PET, due to CT's unique imaging principle based on X-ray absorption. Meanwhile, T2 MRI and cine MRI, which both highlight fluid content, align closely with each other while PET, focused on metabolic activities, shows greater similarity to T2 and cine MRI than to T1 MRI and CT. Further elaboration on these findings is provided in the supplementary. These findings emphasize the capability of Hermes to learn meaningful and accurate prior knowledge from diverse medical images.

Ablation study. We conducted ablation studies using the ResUNet-Small backbone with Hermes to assess the impact of various configurations. We start with Hermes with
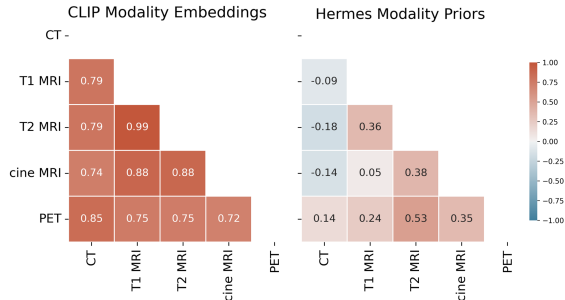
Figure 5. Cosine similarity of Hermes's learned modality priors and CLIP's modality embeddings. The learned modality prior knowledge of Hermes is consistent with imaging principles.

only the task prior applied on the $16\times$ down-sampled scale, which is the most basic configuration of our method. Hierarchically integrating the task priors at multiple scales (e.g., $4\times$, $8\times$, and $16\times$) enhances the average Dice from 81.72% to 82.98%. In the prior fusion module, we employ a bidirectional attention mechanism to concurrently update both the prior and image features. To evaluate the effectiveness of this design, we alter the bidirectional attention to a unidirectional cross-attention. This modification only allows the prior to aggregate information while keeping the image features unchanged, similar to dynamic kernel approaches, e.g. [63, 64]. This change leads to a decreased Dice from 82.98% to 82.04%. Incorporating modality priors further improved the model's capacity to handle complex, heterogeneous multi-modality data, increasing the Dice score from 83.11%. Adding an auxiliary loss can slightly boost the performance to 83.37%.

5. Discussion and Conclusion

Implications of Hermes. Our work introduces a holistic and universal paradigm in medical image analysis, leveraging the diversity and commonality among medical tasks in contrast to developing task-specific models [23, 33]. Despite exhibiting variability, different anatomical structures can share similar visual cues and enhance each other, evidenced by our results on in Table 2. Moreover, the interdisciplinary nature of our paradigm makes it possible to explore into deeper, more complex relationships within medical imagery, as initiated in our analysis (Figs. 4 and 5).

Additionally, the universal paradigm is a cost-efficient approach to scaling medical imaging usage. Given the challenges in assembling large, fully-annotated datasets, Hermes adeptly incorporates diverse, partially annotated datasets, addressing a broad spectrum of clinical targets. This flexibility enables the training of Hermes on a wide range of datasets with varying annotations, distinguishing it from conventional task-specific models and making it also an effective pretraining method with superior transferability. Such adaptability not only makes Hermes a practical asset in medical imaging but also signals a shift towards more versatile, data-inclusive

approaches in the field.

Connection with natural image foundation model. Advances in foundational segmentation models (e.g., Segment-Anything Model (SAM) [38]) have demonstrated notable progress for general vision tasks. Yet general-purpose models struggle to transfer to medical imaging tasks. For instance, a benchmark study [28] comparing SAM against 12 medical imaging tasks revealed that SAM's Dice scores consistently fell behind those of UNet by margins ranging from 0.1 to 0.5, and even reaching up to 0.6-0.7. Efforts of finetuning SAM are considerable to enhance its medical-image performance (e.g. MedSAM [46]). Despite improvements over SAM, MedSAM's performance remains inferior to that of our 3D ResUNet baseline, even when the latter is trained from scratch with a single dataset. For example, on CT liver: ResUNet: 97.64 vs. MedSAM: 91.42, and CT pancreas: ResUNet: 80.45 vs. MedSAM: 76.76. The underperformance of SAM in medical imaging tasks may originate from: (1) the substantial domain gap [49] and (2) its inherent 2D design is ill-equipped to harness the 3D information that is crucial to medical image tasks. Therefore, building a foundation model for 3D medical images should be deeply rooted in 3D medical imaging itself. The proposed universal segmentation paradigm, along with Hermes, offers a versatile means for utilizing diverse, large-scale medical image datasets, opening up perspectives for the development of foundational models in medical imaging.

Limitations. While our model extends to leverages the breadth of partially labeled data, its performance can be potentially improved by addressing missing labels with self-supervised [27] or semi-supervised learning [56, 57]. Integrating more types of medical prior knowledge into model inference presents an intriguing topic for future investigation.

Conclusion. We introduce a universal medical image segmentation paradigm that learns generalizable and transferable representations from diverse medical image tasks, spanning various targets, body regions, and modalities. Following this paradigm, we propose a context-prior learning approach, Hermes, to tackle the challenges of annotation and modality heterogeneity. Hermes can handle multiple tasks by harnessing their synergy via our proposed task priors. Extensive experiments with Hermes underscore the superiority of this universal paradigm in both upstream and downstream tasks. Exploratory analysis of the learned priors shows intriguing relations among tasks and modalities, resonating with the anatomical and imaging principles in radiology.

References

[1] Acgme program requirements for graduate medical education in diagnostic radiology. Online, 2022. Accessed: 2023/11/08. 1

[2] Diagnostic radiology case log categories and required minimum numbers review committee for radiology. Online, 2023. Accessed: 2023/11/08. 1

[3] 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, 6, 7

[4] Isaac Bankman. *Handbook of medical image processing and analysis*. Elsevier, 2008. 1

[5] 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

[6] Mariano Cabezas, Arnau Oliver, Xavier Lladó, Jordi Freixenet, and Meritxell Bach Cuadra. A review of atlas-based segmentation for magnetic resonance brain images. *Computer methods and programs in biomedicine*, 104(3):e158–e177, 2011. 2

[7] 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. Multicentre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. *IEEE Transactions on Medical Imaging*, 40(12):3543–3554, 2021. 5

[8] 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

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 4

[10] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019. 5

[11] Yizhi Chen, Yunhe Gao, Kang Li, Liang Zhao, and Jun Zhao. Vertebrae identification and localization utilizing fully convolutional networks and a hidden markov model. *IEEE Transactions on Medical Imaging*, 39(2):387–399, 2019. 2

[12] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Perpixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 2

[13] 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. 2

[14] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61 (1):38–59, 1995. 2

[15] 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

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

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[18] Yunhe Gao, Rui Huang, Ming Chen, Zhe Wang, Jincheng Deng, Yuanyuan Chen, Yiwei Yang, Jie Zhang, Chanjuan Tao, and Hongsheng Li. Focusnet: Imbalanced large and small organ segmentation with an end-to-end deep neural network for head and neck ct images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*, pages 829–838. Springer, 2019. 1

[19] Yunhe Gao, Chang Liu, and Liang Zhao. Multi-resolution path cnn with deep supervision for intervertebral disc localization and segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 309–317. Springer, 2019. 1

[20] Yunhe Gao, Rui Huang, Yiwei Yang, Jie Zhang, Kainan Shao, Changjuan Tao, Yuanyuan Chen, Dimitris N Metaxas, Hongsheng Li, and Ming Chen. Focusnetv2: Imbalanced large and small organ segmentation with adversarial shape constraint for head and neck ct images. *Medical Image Analysis*, 67: 101831, 2021. 1, 2

[21] Yunhe Gao, Zhiqiang Tang, Mu Zhou, and Dimitris Metaxas. Enabling data diversity: efficient automatic augmentation via regularized adversarial training. In *International Conference on Information Processing in Medical Imaging*, pages 85–97. Springer, 2021. 1

[22] Yunhe Gao, Mu Zhou, and Dimitris N Metaxas. Utnet: a hybrid transformer architecture for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 61–71. Springer, 2021. 1

[23] 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. 4, 6, 8

[24] 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

[25] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part I*, pages 272–284. Springer, 2022. 4, 6

[26] 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. 4, 5

[27] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 8

[28] Sheng He, Rina Bao, Jingpeng Li, P Ellen Grant, and Yangming Ou. Accuracy of segment-anything model (sam) in medical image segmentation tasks. *arXiv preprint arXiv:2304.09324*, 2023. 8

[29] 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

[30] 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

[31] Shah Hussain, Iqra Mubeen, Niamat Ullah, Syed Shahab Ud Din Shah, Bakhtawar Abduljalil Khan, Muhammad Zahoor, Riaz Ullah, Farhat Ali Khan, and Mujeeb A Sultan. Modern diagnostic imaging technique applications and risk factors in the medical field: A review. *BioMed Research International*, 2022, 2022. 1

[32] Juan Eugenio Iglesias and Mert R Sabuncu. Multi-atlas segmentation of biomedical images: a survey. *Medical image analysis*, 24(1):205–219, 2015. 2

[33] 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, 4, 6, 8

[34] 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

[35] 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

[36] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 5

[37] 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

[38] 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. 8

[39] 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

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

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

[42] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4

[43] 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

[44] 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, 5, 7

[45] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer:

Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 5

[46] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023. 8

[47] 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

[48] Ozan Oktay, Enzo Ferrante, Konstantinos Kamnitsas, Mattias Heinrich, Wenjia Bai, Jose Caballero, Stuart A Cook, Antonio De Marvao, Timothy Dawes, Declan P O'Regan, et al. Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. *IEEE transactions on medical imaging*, 37(2):384–395, 2017. 2

[49] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *Advances in neural information processing systems*, 32, 2019. 8

[50] Muhammad Imran Razzak, Saeeda Naz, and Ahmad Zaib. Deep learning for medical image processing: Overview, challenges and the future. *Classification in BioApps: Automation of Decision Making*, pages 323–350, 2018. 1, 2

[51] 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

[52] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 4

[53] David S Sarkany, Anuradha S Shenoy-Bhangle, Tara M Catanzano, Tabitha A Fineberg, Ronald L Eisenberg, and Priscilla J Slanetz. Running a radiology residency program: strategies for success. *Radiographics*, 38(6):1729–1743, 2018. 1

[54] Valerie C Scanlon and Tina Sanders. *Essentials of anatomy and physiology*. FA Davis, 2018. 1

[55] 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

[56] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 8

[57] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 8

[58] 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, 5

[59] Yutong Xie, Jianpeng Zhang, Yong Xia, and Qi Wu. Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In *European Conference on Computer Vision*, pages 558–575. Springer, 2022. 5, 6

[60] Yiwen Ye, Jianpeng Zhang, Ziyang Chen, and Yong Xia. Desd: Self-supervised learning with deep self-distillation for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 545–555. Springer, 2022. 5, 6, 7

[61] 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, 5, 7

[62] 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. 4

[63] 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, 5, 7, 8

[64] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 34:10326–10338, 2021. 2, 8

[65] 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

[66] 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