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

Supplementary Material

1. Supplement method

1.1. Prior fusion module details

The prior fusion module is an essential part of Hermes, which is responsible for integrating knowledge encapsulated in prior tokens with image feature maps. This fusion process utilizes attention modules, enabling efficient and global information exchange between prior tokens and feature maps. This design ensures Hermes' compatibility with existing backbones. We illustrate several implementations for common backbones in Figure 1. We first introduce the formulation with the conventional attention module. Then, we introduce the bi-directional cross-attention for CNN backbones. Finally, we show how to merge the prior fusion into the MedFormer backbone.

Conventional attention module. Conventional Transformer architectures (e.g. ViT [5]) often rely on multi-head self-attention that captures the all-to-all pairwise dependencies among input tokens. To extend, we present a general formulation of merging prior fusion into a conventional attention module, see Figure 1 (C). Given the input feature map $X \in \mathcal{R}^{n \times C}$ ($n = D \times H \times W$ is the number of tokens in the feature map, C denotes the token dimension), and the prior tokens $\boldsymbol{p} \in \mathcal{R}^{(|\boldsymbol{t}_k|+l) \times C}$, where $|\boldsymbol{t}_k|$ denotes the number of tasks, and l is the length of each modality prior. We first concatenate the feature map and the prior tokens together as the input of the transformer block: $I = [X, \boldsymbol{p}] \in \mathcal{R}^{n+|\boldsymbol{t}_k|+l}$. The conventional Transformer block operates as follows:

$$
\begin{aligned}
I' &= \text{LN}(I) \\
I'' &= \text{MHA}(I', I', I')) + I \\
\text{MHA}(q, k, v) &= [head_1, \cdots, head_h] W^O \\
head_i &= \text{Attention}(qW_i^Q, kW_i^K, vW_i^V) \\
\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\intercal}{\sqrt{d_h}})\mathbf{V} \\
\hat{I} &= \text{FFN}(\text{LN}(I'')) + I''
\end{aligned}
\tag{1}
$$

where W_Q, W^K, W^V, W^O are weight matrices, and d_h is the dimension of each head. We can then obtain the prior-injected feature map and the posterior tokens by splitting $\hat{I} = [\hat{X}, \hat{\boldsymbol{p}}]$. The computation complexity for the above prior-modified attention module is $O((n + |\boldsymbol{t}_k| + l)^2)$

CNN backbones. CNN backbones, like ResUNet [8, 18], usually use hierarchical architectures. The quadratic complexity of the above all-to-all attention makes it unaffordable to apply the prior fusion module on high-resolution feature maps (e.g. $n = 32,768$ for a $32 \times 32 \times 32$ feature map). Therefore, we implement the prior fusion module with a bi-directional cross-attention module instead, see Figure 1 (A). The prior tokens $\boldsymbol{p}$ first aggregate image-specific information from the feature map X to obtain the posterior tokens $\hat{\boldsymbol{p}}$:

$$
\begin{aligned}
\boldsymbol{p}' &= \text{LN}(\boldsymbol{p}) \\
X' &= \text{LN}(X) \\
\boldsymbol{p}'' &= \text{MHA}(\boldsymbol{p}', X', X') + \boldsymbol{p} \\
\hat{\boldsymbol{p}} &= \text{FFN}(\text{LN}(\boldsymbol{p}'')) + \boldsymbol{p}''
\end{aligned}
\tag{2}
$$

Then we inject the knowledge in the posterior tokens to obtain the prior-injected feature map $\hat{X}$:

$$
\begin{aligned}
\hat{\boldsymbol{p}}' &= \text{LN}(\hat{\boldsymbol{p}}) \\
X'' &= \text{MHA}(X', \hat{\boldsymbol{p}}', \hat{\boldsymbol{p}}') + X \\
\hat{X} &= \text{FFN}(\text{LN}(X'')) + X''
\end{aligned}
\tag{3}
$$

As $|\boldsymbol{t}_k| + l \ll n$, the computation complexity of the bi-direction cross-attention for the prior fusion module is $O(n)$. For example, for the dataset with the most tasks, AMOS CT, $|\boldsymbol{t}_k| + l = 25 \ll n = 32,768$ for a $32 \times 32 \times 32$ feature map. With this design, the prior fusion module can adaptively integrate the prior tokens and the feature maps with minor additional computational costs. We implement Hermes-R by inserting the cross-attention module at the end of each stage of the ResUNet, i.e. after the convolution layers at $4\times$, $8\times$, and $16\times$ downsampling scales.

MedFormer. MedFormer [6] is a Transformer model proposed for medical image segmentation. One key component of MedFormer is its B-MHA attention module, which incorporates a compressed semantic map to reduce computation complexity as well as enhance representation learning. In Figure 1 (B), we show our implementation of merging the prior fusion module into the B-MHA module of the MedFormer backbone. In B-MHA, M is a semantic map that is encoded and refined for rich semantic information within a much lower spatial resolution compared to the feature map X. We concatenate the semantic map M with the prior tokens $\boldsymbol{p}$: $I_M = [M, \boldsymbol{p}]$. The X and I_M are linearly projected to $\mathbf{Q}/\mathbf{K}/\mathbf{V}$ and $\bar{\mathbf{Q}}/\bar{\mathbf{K}}/\bar{\mathbf{V}}$ respectively. To reduce the computation, B-MHA shares the query and key of X and I_M, i.e. $\mathbf{Q} = \mathbf{K}$ and $\bar{\mathbf{Q}} = \bar{\mathbf{K}}$, as the dot product of the query and key in cross-attention measures the similarity of token pairs of two inputs, which is symmetrical. The attention matrix is reused by simply transposing the dot product matrix:

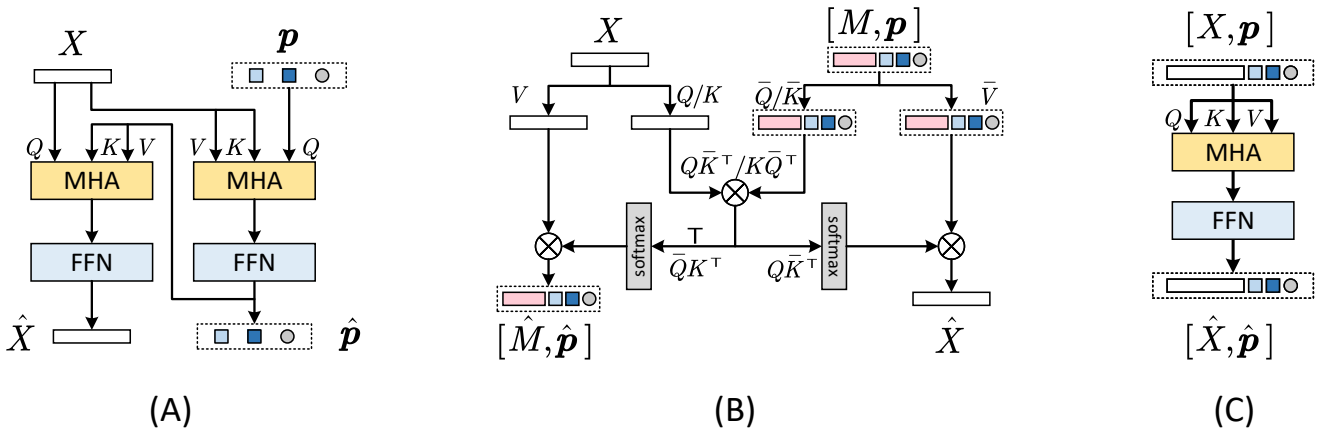

Figure 1. The implementation of prior fusion module for existing backbones. (A) The implementation for CNN backbones, like ResUNet [18]. We use a bi-directional cross-attention module to process both the feature map X and the prior tokens $\boldsymbol{p}$. (B) The implementation for MedFormer [6]. We merge the prior tokens into the semantic map of its B-MHA module. (C) The implementation for conventional attention module, e.g. ViT [5] backbone. The normalization layers and residue connections for all three implementations are omitted for simplicity.

$$\begin{aligned}\hat{X} &= \text{Attention}(\mathbf{Q}, \bar{\mathbf{K}}, \bar{\mathbf{V}}) = \text{softmax}(\frac{\mathbf{Q}\bar{\mathbf{K}}^{\mathsf{T}}}{\sqrt{d}})\bar{\mathbf{V}} \\ \hat{I}_M &= \text{Attention}(\bar{\mathbf{Q}}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\bar{\mathbf{Q}}\mathbf{K}^{\mathsf{T}}}{\sqrt{d}})\mathbf{V} \\ (\mathbf{Q}\bar{\mathbf{K}}^{\mathsf{T}})^{\mathsf{T}} &= (\mathbf{K}\bar{\mathbf{Q}}^{\mathsf{T}})^{\mathsf{T}} = \bar{\mathbf{Q}}\mathbf{K}^{\mathsf{T}}\end{aligned} \tag{4}$$

The normalization layer, FFN, and residue connections are not included in the equation for simplicity. More details can be found in the original MedFormer paper. We can then obtain the updated semantic map and the posterior tokens by splitting $\hat{I}_M = [\hat{M}, \hat{\boldsymbol{p}}]$. As $|M| + |\boldsymbol{t}_k| + l \ll n$, the computation complexity is $O(n)$. We implement Hermes-M by incorporating the prior fusion module with the B-MHA module on the $4\times$, $8\times$ and $16\times$ downsampling scales within the MedFormer backbone.

Computation comparison. In Table 1, we present the GPU memory usage, number of parameters, and inference time for Hermes with various backbones. For the ResUNet backbone, thanks to the efficient bi-directional cross-attention, Hermes-R only slightly increases GPU memory usage and inference time, despite additional parameters due to the cross-attention in the prior fusion module. For the Transformer backbone MedFormer, Hermes-M demonstrates almost identical consumption on memory, inference time, and the number of parameters compared with MedFormer. These results exemplify the efficacy of integrating the prior fusion module into MedFormer's existing attention module, highlighting Hermes' ability in leveraging different backbones without significantly impacting the required computational resources.

Table 1. Computation comparison between the Hermes and the corresponding backbone. The memory consumption and inference time are measured with an image size of $2 \times 1 \times 128 \times 128 \times 128$ on one Nvidia A100 GPU. We report the average inference time over 100 runs.

Model	Memory/G	#Params/M	Inference Time/s
ResUNet	11.23	40.56	0.13
Hermes-R	11.54	59.61	0.16
MedFormer	11.44	43.20	0.19
Hermes-M	11.62	44.50	0.20

2. Supplement Experiments

Visual analysis on posterior prototypes. In Figure 2, we show the heatmap visualization of each posterior prototype with the output feature map of the decoder measured with dot-product similarity. The visualization underscores the quality of Hermes's predicted posterior prototypes, which adeptly capture the semantic essence of each category, aligning closely with the respective feature maps. Remarkably, Hermes manages precise predictions even for small organs with complex shapes. Take the right and left adrenal glands as an example: despite their tiny size and irregular shape, the posterior prototypes predicted by Hermes accurately reflect their intricate edges.

Comparison with other methods. We provide the detailed performance of other comparison methods on each dataset, see Table 2. The ResUNet is trained under the traditional paradigm. All comparison methods are trained with the universal paradigm and are implemented with ResUNet for a fair comparison. All methods under the universal paradigm exhibit better performance compared with tradi-

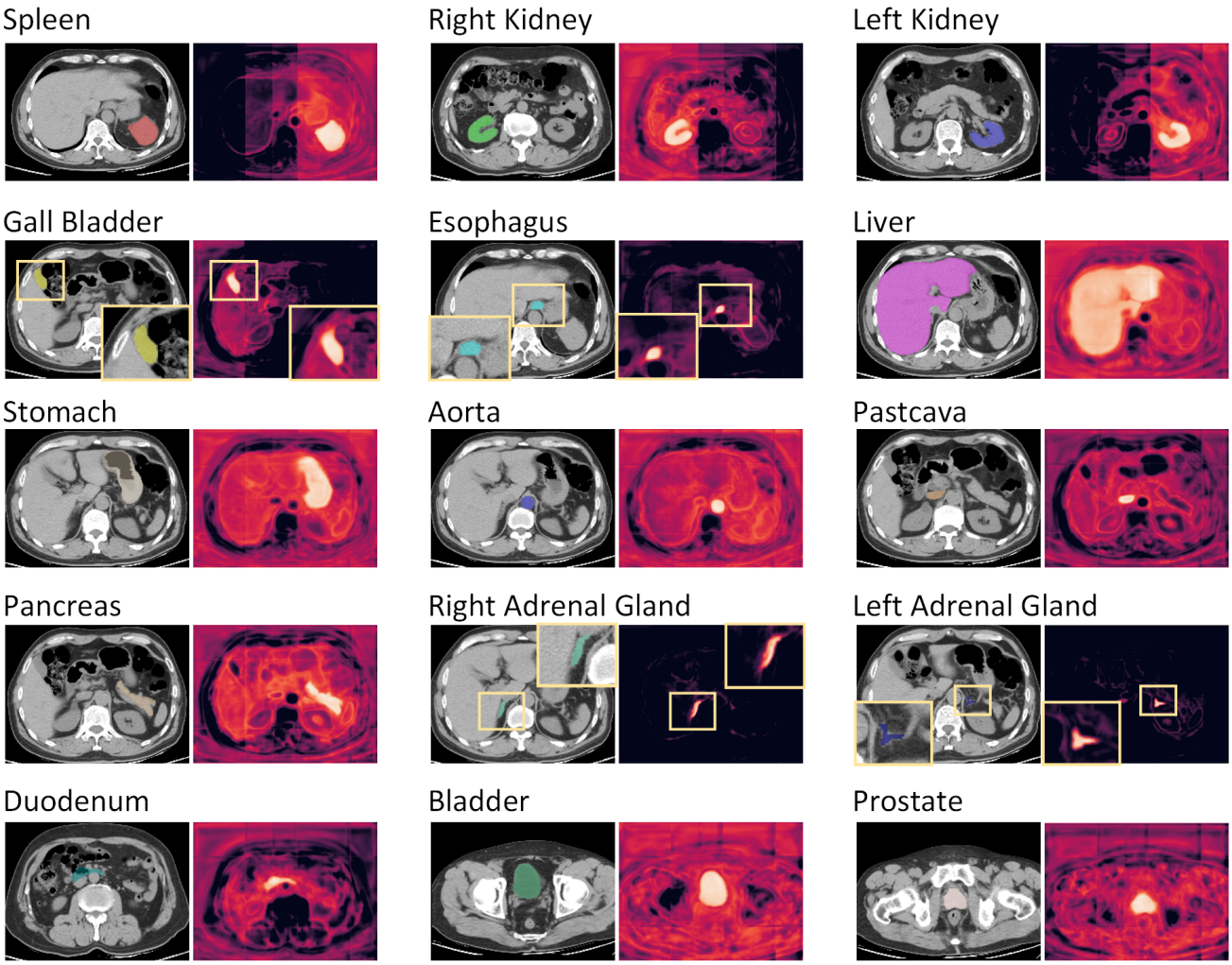

Figure 2. Heatmap visualization of each posterior prototype for the AMOS CT dataset. The brighter means the higher similarity. The yellow boxes are the zoomed-in version of small organs for better visualization.

tionally trained ResUNet. Among universal method settings, our Hermes-R shows a consistent advantage on the eleven upstream datasets.

The detailed number of Figure 3 (A) in the main text is presented in Table 3. All methods under the universal paradigm show better than the ResUNet trained on each individual dataset. Hermes demonstrates consistent advantages, especially on difficult classes, modality PET, tumor&lesion classes, and head&neck region tasks.

Ablation on the length of modality prior token. We present an additional ablation study on the length of the modality prior token, see Table 4. We follow the setting in the ablation study section in the manuscript using the ResUNet-Small backbone to implement Hermes. Given that each modality encompasses a significant amount of variation, we find that a modality prior of length 1 does not possess adequate capacity to encode modality-related information. We observe that longer modality tokens further improve the performance, but the gains saturate as the length increases. Therefore, we choose $l = 10$ for our main experiments.

Additional analysis on the learned priors. In this section, we briefly introduce the imaging principle and the typical visual appearance of structures in different modalities to help interpret Hermes's learned modality priors, see Fig. 3.

CT uses X-ray beams to create detailed cross-sectional images of the body. In CT images, bones and other dense structures appear very bright (high attenuation), while soft tissues show up in varying shades of gray. Air and other gas-filled spaces appear dark due to their low X-ray absorption. CT is particularly effective for visualizing bones, lung tissue, and detecting abnormalities like tumors or fractures.

T1-weighted MRI utilizes magnetic fields and radiofrequency pulses to produce detailed images of the body's internal structures. In T1-weighted images, fat-containing tissues

Table 2. Comparison with other methods. ResUNet is trained with the traditional paradigm, while all comparison methods are reimplemented with the ResUNet backbone for fair comparison and extend to the universal medical image segmentation paradigm.

Model	BCV	SS T	SS H	LiTS T	KiTS T	AMOS CT	AMOS MR	CHAOS	M&Ms	AutoPET	DLBS	AVG
ResUNet	84.36	88.59	78.12	64.87	81.89	88.97	85.43	91.34	85.73	65.52	94.31	82.65
Multi-decoder [4]	83.90	89.18	78.31	65.74	81.66	89.27	85.65	91.56	86.00	66.06	94.71	82.91
DoDNet [21]	85.02	88.87	78.49	65.84	82.65	88.86	86.22	91.35	85.97	67.49	94.94	83.25
CLIP-driven [15]	85.12	89.34	78.50	65.37	82.83	88.94	86.39	91.81	86.04	66.78	95.17	83.30
UniSeg [20]	85.32	89.39	78.69	65.80	82.96	89.17	86.55	91.85	86.26	70.12	95.34	83.77
MultiTalent [19]	85.18	89.18	80.01	65.33	82.25	89.13	86.57	91.55	86.28	71.51	95.75	83.88
Hermes-R	**85.99**	**89.50**	**80.62**	**67.49**	**85.53**	**89.63**	**86.78**	**92.01**	**86.94**	**73.69**	**96.21**	**84.95**

Table 3. The detailed number of Figure 3 (A) in the main text. We compare traditionally trained ResUNet and other SOTA method under the universal paradigm in six aspects. "Difficult Classes" are the classes that have Dice scores lower than 80 under the traditional paradigm.

Model	Difficult Classes	Modality PET	Tumor&Lesion	Region Head&Neck	All Classes	All Datasets
ResUNet	70.67	65.52	70.76	78.13	84.54	82.65
Multi-decoder	71.42	66.06	71.15	78.31	84.70	82.91
DoDNet	72.17	67.49	71.99	78.48	84.93	83.25
CLIP-driven	72.11	66.78	71.64	78.50	85.05	83.30
UniSeg	72.46	70.12	72.96	78.69	85.27	83.77
MultiTalent	73.18	71.51	73.03	80.01	85.57	83.88
Hermes-R	74.35	73.69	75.57	80.62	86.16	84.95

Table 4. Ablation study on the length of modality prior token. We report the average Dice on the seven datasets with Hermes implemented with the ResUNet-Small backbone.

l	1	5	10	20
Avg Dice	82.93	83.16	83.37	83.38

appear bright, and water-rich tissues look darker. This contrast makes T1 MRI particularly useful for visualizing fine anatomical details, such as the brain's white and gray matter.

T2-weighted MRI also uses magnetic fields and radio waves but with different timing parameters than T1, leading to different tissue contrasts. In T2 images, fluid-containing tissues appear bright, while fat appears darker compared to T1 images. This characteristic makes T2 MRI ideal for visualizing fluid-filled spaces and edema.

Cine MRI is a specialized form of MRI used to capture moving images of the body. It is particularly useful in cardiology for visualizing the heart's movement and blood flow. In cine MRI, fluid dynamics are emphasized, making it excellent for assessing cardiac function, valve abnormalities, and congenital heart disease. The visualization of blood flow and moving structures is a unique aspect of cine MRI.

PET (Positron Emission Tomography) scans use radioactive tracers to detect metabolic activities within the body. The patient is injected with a radiotracer, which accumulates in areas of high metabolic activity. PET scanners detect the gamma rays emitted by the tracer and use this data to construct images. In PET images, areas of high tracer uptake, such as rapidly growing cancer cells, appear brighter. PET is highly effective in cancer diagnosis, as it can reveal the metabolic activity of tumors.

In Fig. 3, we present the cosine similarity between the modality priors as learned by Hermes. Consistent with imaging principles, Hermes identifies CT as distinctly different from MRI and PET modalities, owing to its unique X-ray based imaging technique and contrasting tissue visualization. This finding aligns with the principle that CT images bone and dense structures more effectively, setting it apart from MRI and PET techniques.

Regarding MRI sequences, Hermes notes a higher similarity among them compared to CT, reflecting their shared basis in magnetic resonance techniques, even though they provide different tissue contrasts. Specifically, Hermes discerns a closer relationship between cine MRI and T2 MRI, attributable to their shared emphasis on fluid content visualization. This is in line with cine MRI's application in capturing moving structures like blood flow, similar to the fluid-highlighting characteristics of T2 MRI images.

Additionally, Hermes finds PET imaging to be more similar to T2 and cine MRI than to T1 MRI. This observation can be understood through PET's focus on metabolic activities and functional changes, aspects that are also emphasized to some extent in T2 and cine MRI, despite their different fundamental principles. On the other hand, the lower similarity with T1 MRI is logical, given T1 MRI's distinct imaging focus, primarily on fat visualization, as opposed to PET's metabolic activity emphasis.

These findings by Hermes are in accordance with the fun-

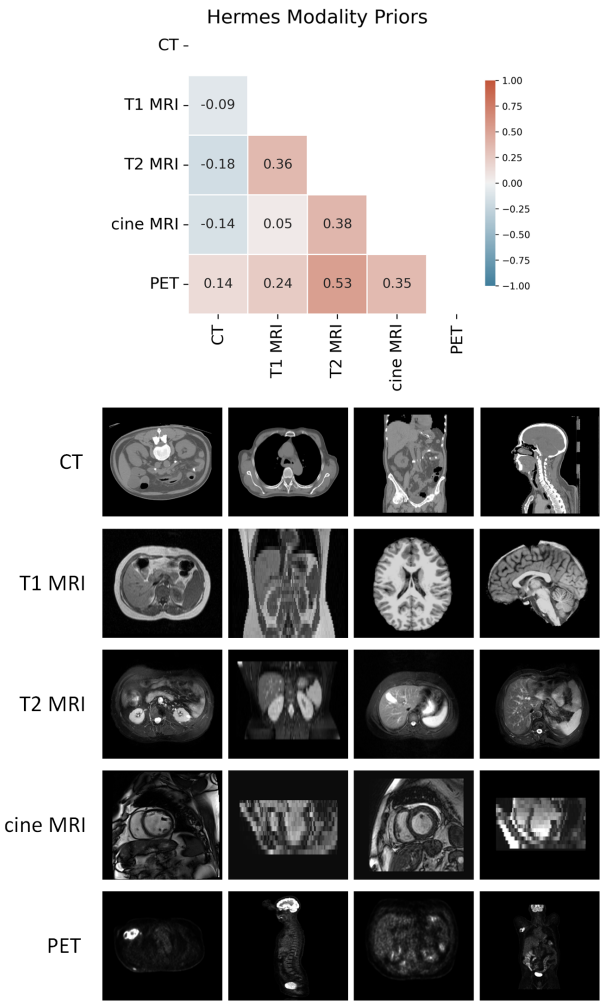

Figure 3. Upper: The cosine similarity between the learned Hermes modality priors. Lower: Illustration of each imaging modality.

damental principles of these imaging modalities. They indicate that Hermes is effectively capturing the unique imaging characteristics and tissue contrasts inherent to each modality, demonstrating a sophisticated understanding of how different imaging techniques visualize various tissues and physiological processes.

3. Dataset details

In this section, we provide detailed information about each dataset, including the volume of data, annotation categories, data sources, as well as how we use them for training and testing. At last, we explain how we designed our experiments using these datasets.

BCV dataset. The BCV dataset [13] (Multi-Atlas Labeling Beyond the Cranial Vault) comprises 50 subjects with abdominal CT scans, of which 30 training images are publicly available. Thirteen abdominal organs were manually

Table 5. 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 [13]	Abdomen	CT	Organs	13	30
LiTS [2]	Abdomen	CT	Liver & Tumor	2	131
KiTS [9]	Abdomen	CT	Kidney & Tumor	2	210
AMOS CT [10]	Abdomen	CT	Organs	15	300
SS T[14]	Thorax	CT	Organs	6	50
SS H [14]	Head & Neck	CT	Organs	22	50
AMOS MR [10]	Abdomen	MRI	Organs	13	60
CHAOS [11]	Abdomen	T1 & T2 MRI	Organs	4	60
M&Ms [3]	Cardiac	cineMRI	Structures	3	320
DLBS [17]	Brain	T1 MRI	Structures	3	213
AutoPET [7]	Whole body	PET	Lesions	1	1014
SegTHOR [12]	Thorax	CT	Organs	3	40
MSD Pancreas [1]	Abdomen	CT	Pancreas & Tumor	2	281

labeled on a volumetric basis using the MIPAV software. The labeled organs include the spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, portal vein and splenic vein, pancreas, right adrenal gland, and left adrenal gland. Some patients may lack the right kidney or gallbladder, and therefore these organs are not labeled. All scans were acquired for routine clinical care from CT scanners at the Vanderbilt University Medical Center (VUMC). The BCV dataset is used as one of the seven datasets for upstream training. We randomly select 75%/5%/20% of the publicly available images for training/validation/testing.

LiTS dataset. The LiTS dataset [2] (Liver Tumor Segmentation Challenge) comprises 201 computed tomography (CT) images of the abdomen, with 131 training cases and 70 testing cases, where only the label of training cases are publicly available. The LiTS dataset provides detailed annotation for tumors while offering coarse annotation for the liver. The image data originates from various clinical sites, including Ludwig Maxmilian University of Munich, Radboud University Medical Center of Nijmegen, Polytechnique & CHUM Research Center Montréal, Tel Aviv University, Sheba Medical Center, IRCAD Institute Strasbourg, and the Hebrew University of Jerusalem. The studied subjects suffer from diverse liver tumor diseases, such as hepatocellular carcinoma (HCC), as well as secondary liver tumors and metastases originating from colorectal, breast, and lung cancers. The tumors exhibit varying contrast enhancement, including hyper and hypo-dense contrast. The images represent a mix of pre- and post-therapy abdominal CT scans, acquired with different CT scanners and acquisition protocols. The LiTS dataset is used as one of the seven datasets for upstream training. We randomly select 75%/5%/20% of the 131 training cases for training/validation/testing.

KiTS dataset. The KiTS19 dataset [9] comprises segmented CT imaging and treatment outcomes for 300 patients who underwent partial or radical nephrectomy for one or

more kidney tumors at the University of Minnesota Medical Center between 2010 and 2018. Out of these cases, 210 have been released publicly, while the remaining 90 are kept private for evaluation purposes. The KiTS is used as one of the seven datasets for upstream training. We randomly select 75%/5%/20% of the 210 training cases for training/validation/testing.

AMOS CT & MR dataset. The AMOS dataset [10] is a large-scale collection of CT and MRI data from 600 patients diagnosed with abdominal tumors or abnormalities at Longgang District People's Hospital. The dataset comprises 500 CT and 100 MRI scans acquired from eight different scanners and vendors, encompassing 15 organ categories: spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, pancreas, right adrenal gland, left adrenal gland, duodenum, bladder, and prostate/uterus. For CT images, AMOS provides 200 scans for training and 100 scans for validation, while for MRI images, 40 scans are provided for training and 20 scans for validation. Both AMOS CT and AMOS MR are used as two of the seven datasets for upstream training. In line with the AMOS benchmark paper [10], we report testing performance on the official validation set and utilize 95%/5% training data for model training/validation. As all images in the AMOS MR validation set don't have the annotation of bladder and prostate, we only segment 13 organs for AMOS MR.

StructSeg dataset. The StructSeg dataset [14] is collected from a challenge for the segmentation of organs-at-risk (OAR) and gross target volume (GTV) of tumors of two types of cancers, nasopharynx cancer and lung cancer, for radiation therapy planning. We use Task 1 and 3, organ-at-risk segmentation from head&neck and thorax CT scans in our experiments, denoted as SS H and SS T respectively. SS H has 22 OAR annotations from 50 nasopharynx cancer patients, including left eye, right eye, left lens, right lens, left optical nerve, right optical nerve, optical chiasma, pituitary, brain stem, left temporal lobes, right temporal lobes, spinal cord, left parotid gland, right parotid gland, left inner ear, right inner ear, left middle ear, right middle ear, left temporomandibular joint, right temporomandibular joint, left mandible and right mandible. SS T has 6 OARs annotated on CT scans from 50 lung cancer patients, including left lung, right lung, spinal cord, esophagus, heart, and trachea. We split the scans into 75%/5%/20% for training/validation/testing.

CHAOS dataset. The CHAOS dataset [11] is collected from a challenge for the precise segmentation of abdominal organs. We use the data from Task 5: segmentation of abdominal organs from MRI. Four organs, including the liver, left kidney, right kidney, and spleen are annotated. They provide three MR sequences, including T1-in-phase, T1-out-phase, and T2-SPIR, for 20 patients. We treat different MR sequences as separate images and split the dataset at the patient level into 75%/5%/20% for training/validation/testing.

M&Ms dataset. The M&Ms dataset [3] is from Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation (M&Ms) Challenge, which was organized as part of the MICCAI 2020 Conference. This dataset cohort includes patients with hypertrophic and dilated cardiomyopathies and healthy subjects. All subjects were scanned in clinical centres in three different countries (Spain, Germany, and Canada) using four different MRI scanner vendors (Siemens, General Electric, Philips, and Canon). The training set contains 150 annotated images from two vendors (75 each), while the testing set contains 170 cases (20 for the first vendor and 50 each for the other three vendors). Three categories of annotation are 61 provided at the end-diastolic (ED) and end-systolic (ES) 62 phase, including left ventricle (LV), right ventricle (RV), and 63 left ventricular myocardium (MYO). We use the official testing set for testing, and divide the training set into 95% for training and 5% for validation.

DLBS dataset. The Dallas Lifespan Brain Study (DLBS) [17] is designed to understand the antecedents of preservation and decline of cognitive function at different stages of the adult lifespan, with a particular interest in the early stages of a healthy brain's march towards Alzheimer Disease. We use the 213 T1 MRI scans to segment the cerebrospinal fluid, gray matter and white matter. Following [16], we divide the 213 scans into 129 for training, 43 for validation, and 43 for testing.

AutoPET dataset. The AutoPET dataset [7] provides annotated Positron Emission Tomography/Computed Tomography (PET/CT) studies, encompassing a significant collection of 1014 whole-body Fluorodeoxyglucose (FDG)-PET/CT datasets. This dataset includes 501 studies from patients diagnosed with malignant lymphoma, melanoma, and non-small cell lung cancer (NSCLC), alongside 513 studies serving as negative controls without PET-positive malignant lesions. We divide the dataset at the patient level into 75%/5%/20% for train/validation/testing.

SegTHOR dataset (SS T and SS H). The SegTHOR dataset [12] aims at the thoracic organ-at-risk segmentation in CT images. This dataset provides 4 OARs annotations from 40 CT scans, including heart, aorta, trachea, and esophagus. We use the SegTHOR dataset as a downstream task to evaluate the generalization of models. We directly use the upstream-trained model to make predictions on all 40 images and report the generalization performance.

MSD pancreas & tumor dataset. The MSD pancreas & tumor dataset is a part of the Medical Image Segmentation Decathlon (MSD) [1], an international challenge aimed at identifying a general-purpose algorithm for medical image segmentation. The competition encompasses ten distinct datasets featuring various target regions, modalities, and challenging attributes. MSD pancreas & tumor is one of the datasets that is annotated for pancreas and tumors. The

shape and position of tumors vary greatly between patients. The MSD pancreas & tumor dataset consists of 281 CT images. We use it as a downstream task to evaluate models' capacity for transfer learning and incremental learning. We split the dataset into 214 samples for training, 10 samples for validation, and 57 samples for testing. To evaluate the impact of downstream data volume, we conducted experiments on 1%, 10%, 50%, and 100% of the 214 training samples. To reduce the variability from the training sample selection, we report the average performance over 5 runs for the 1% and 10% settings.

Experiment design. To substantiate the efficacy of the proposed universal medical image segmentation paradigm, we have meticulously curated these datasets, see Table 5. These datasets were selected based on three main factors: anatomical regions, imaging modalities, and clinical targets. The careful selection of these upstream training datasets is designed to provide comprehensive answers to the three research questions originally posed in our introduction section. For the downstream tasks, we chose the challenging MSD pancreas & tumor dataset for transfer learning and incremental learning. The pancreas is a relatively small, elongated glandular organ, while the shape and location of a pancreatic tumor can greatly vary. As such, the segmentation difficulty of this task is extremely high. Furthermore, this dataset is comprised of a large number of images, with 281 CT scans, allowing us to adequately test the model's transfer learning and incremental learning abilities under various downstream data volumes. In addition, we select the SegTHOR dataset to verify the model's generalization performance. There is only one thoracic dataset (StructSeg) in the upstream training. The StructSeg and SegTHOR are both for thoracic OAR segmentation and have three overlap categories of heart, trachea, and esophagus. Evaluating performance on these overlapping categories allows us to explore the universal paradigm's potential generalization ability to different anatomical regions and analyze whether more abdominal tasks contribute positively to the generalization of thoracic tasks.

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, 6

[2] 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

[3] 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, 6

[4] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019. 4

[5] 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. 1, 2

[6] 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, 2

[7] 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, 6

[8] 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. 1

[9] 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

[10] 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, 6

[11] 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, 6

[12] 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, 6

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

[14] Hongsheng Li, Jinghao Zhou, Jincheng Deng, and Ming Chen. Automatic structure segmentation for radiotherapy planning challenge 2019. 5, 6

[15] 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. 4

[16] Vishwanatha M Rao, Zihan Wan, Soroush Arabshahi, David J Ma, Pin-Yu Lee, Ye Tian, Xuzhe Zhang, Andrew F Laine, and Jia Guo. Improving across-dataset brain tissue segmentation for mri imaging using transformer. *Frontiers in Neuroimaging*, 1:1023481, 2022. 6

[17] 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, 6

[18] 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. 1, 2

[19] 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. 4

[20] 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. 4

[21] 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. 4