Medical Image Analysis 97 (2024) 103226

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Universal and extensible language-vision models for organ segmentation and tumor detection from abdominal computed tomography

Jie Liu [a], Yixiao Zhang [b], Kang Wang [c], Mehmet Can Yavuz [c], Xiaoxi Chen [d], Yixuan Yuan [e], Haoliang Li [a], Yang Yang [c], Alan Yuille [b], Yucheng Tang [f], Zongwei Zhou [b,*]

[a] City University of Hong Kong, Hong Kong
[b] Johns Hopkins University, United States of America
[c] University of California, San Francisco, United States of America
[d] University of Illinois Urbana-Champaign, United States of America
[e] Chinese University of Hong Kong, Hong Kong
[f] NVIDIA, United States of America

ARTICLE INFO

MSC:
41A05
41A10
65D05
65D17

Keywords:
Organ segmentation
Tumor detection
Vision-language models

ABSTRACT

The advancement of artificial intelligence (AI) for organ segmentation and tumor detection is propelled by the growing availability of computed tomography (CT) datasets with detailed, per-voxel annotations. However, these AI models often struggle with flexibility for partially annotated datasets and extensibility for new classes due to limitations in the one-hot encoding, architectural design, and learning scheme. To overcome these limitations, we propose a universal, extensible framework enabling a single model, termed Universal Model, to deal with multiple public datasets and adapt to new classes (e.g., organs/tumors). Firstly, we introduce a novel language-driven parameter generator that leverages language embeddings from large language models, enriching semantic encoding compared with one-hot encoding. Secondly, the conventional output layers are replaced with lightweight, class-specific heads, allowing Universal Model to simultaneously segment 25 organs and six types of tumors and ease the addition of new classes. We train our Universal Model on 3410 CT volumes assembled from 14 publicly available datasets and then test it on 6173 CT volumes from four external datasets. Universal Model achieves first place on six CT tasks in the Medical Segmentation Decathlon (MSD) public leaderboard and leading performance on the Beyond The Cranial Vault (BTCV) dataset. In summary, Universal Model exhibits remarkable computational efficiency (6× faster than other dataset-specific models), demonstrates strong generalization across different hospitals, transfers well to numerous downstream tasks, and more importantly, facilitates the extensibility to new classes while alleviating the catastrophic forgetting of previously learned classes. Codes, models, and datasets are available at https://github.com/ljwztc/CLIP-Driven-Universal-Model.

1. Introduction

Computer tomography (CT) is a widely used and powerful tool for disease diagnosis and treatment planning (Mattikalli et al., 2022; Zhou et al., 2022; Qu et al., 2023; Chen et al., 2024a). In routine clinical workflow, radiologists analyze hundreds of 2D slices in a single CT volume to find and interpret diagnostic information, which can be tedious and prone to misdiagnosis (Zhou, 2021). Medical image segmentation offers a promising solution, improving diagnostic efficiency and quality by automatically identifying organs, delineating their boundaries, and highlighting abnormalities (Liu et al., 2021; Hu et al., 2023; Chen et al., 2023a, 2024b; Lai et al., 2024).

The progress of medical image segmentation relies heavily on specialized datasets. These include organ/tumor-specific datasets, e.g., LiTS (Liver Tumor Segmentation) (Bilic et al., 2019), KiTS (Kidney Tumor Segmentation) (Heller et al., 2019), and MSD (Medical Segmentation Decathlon) (Simpson et al., 2019), to abdominal multi-organ annotation datasets like BTCV (Beyond The Cranial Vault) (Landman et al., 2015), AMOS (Abdominal Multi-organ Segmentation) (Ji et al., 2022), and AbdomenAtlas (Li et al., 2024). Additionally, Wasserthal et al. (2023) have proposed to provide a holistic anatomical view of the whole body instead of the specific body region, which aims to capture the comprehensive anatomical structure of the human body.

* Corresponding author.
E-mail addresses: yuchengt@nvidia.com (Y. Tang), zzhou82@jh.edu (Z. Zhou).

https://doi.org/10.1016/j.media.2024.103226
Received 20 December 2023; Received in revised form 30 March 2024; Accepted 27 May 2024
Available online 4 June 2024

Considering the highly complex nature of anatomical structures in the human body and the need for fine-grained clinical requirements, we can expect the emergence of new organ/tumor annotations (Jaus et al., 2023), such as vermiform appendix and spleen tumors. Moreover, there is a growing demand for more detailed anatomy annotations, such as distinguishing between the right lobe liver and left lobe liver (Germain et al., 2014). While liver annotation is common in current datasets, the partition of liver lobes remains under-investigated (Bilic et al., 2023), leading to potential overlapping annotations between liver lobes and overall liver in the future. The latest endeavors to meet these emerging requirements involve re-annotating existing datasets with human-in-the-loop and retraining models accordingly (Qu et al., 2023; Wasserthal et al., 2023; Jaus et al., 2023). However, this approach incurs significant annotation costs, particularly for 3D medical imaging, as well as substantial computational consumption for retraining models from scratch (Zhang et al., 2023b, 2024). Therefore, it is important to explore a new framework that can effectively handle new organ/tumor annotations while mitigating the computational burden associated with retraining models.

Several challenges must be addressed when designing such a continual multi-organ segmentation and tumor detection framework. The first challenge involves the encoding orthogonality problem. Existing methods commonly employ one-hot encoding, ignoring the semantic relationship between organs and tumors. For instance, when utilizing one-hot encoding for liver [1,0,0], pancreas [0,1,0], and pancreas tumor [0,0,1], there is no semantic distinction among these classes. A possible solution is the few-hot encoding (Shi et al., 2021), where the liver, pancreas, and pancreas tumor can be encoded as [1,0,0], [0,1,0], and [0,1,1], respectively. Although few-hot labels could indicate that pancreatic tumors are part of the pancreas, the relationship between organs remains orthogonal. In addition, these encoding approaches become challenging to extend when adding new organs into the current framework, as the entire encoding scheme needs to be modified to accommodate the new classes.

The second challenge lies in the label taxonomy inconsistency, where publicly available datasets usually contain annotations of different classes. This disparity in label spaces restricts the traditional segmentation head, which employs a pre-defined kernel number and Softmax activation function, to produce single-class predictions for each voxel based on fixed classes (Liu et al., 2022a; Ma et al., 2021; Schoppe et al., 2020). Consequently, it lacks the flexibility to accommodate new organ/tumor annotations and handle detailed anatomy annotations. For example, "Right Lobe Liver" is part of the "Liver" and "Kidney Tumor" represents a sub-volume of the "Kidney", which means that certain voxels do not exclusively belong to a single specific class as per clinical requirements. One trivial solution is to replace the segmentation head and retrain from scratch, but this would forget the learned knowledge and cause computational consumption. This motivates us to design a novel framework for segmentation heads that allows for flexible prediction and extensibility to new classes.

The third challenge pertains to the design of a dynamic network architecture for continual learning. When it comes to annotating new organs/tumors, the medical segmentation model should be able to extend its segmentation capabilities without accessing previous data, in order to comply with medical privacy regulations. Existing works typically freeze the encoder and decoder, and add an additional decoder when learning new classes (Ji et al., 2023), leading to tremendous memory costs for network parameters. Furthermore, the newly added decoder lacks awareness of the relationship between the newly added organs/tumors and existing classes, resulting in limited utilization of existing knowledge. Hence, our objective is to design a network architecture that minimizes the addition of parameters during different continual learning steps and ensures awareness of the relationship between newly added organs/tumors and existing classes.

To overcome these challenges, we propose a universal and extensible framework that offers a promising solution to multi-organ segmentation and tumor detection. Firstly, we propose the language-driven parameter generator (LPG) to employ the language embeddings, learned from large language models, as a novel semantic encoding method. This enables the capture of semantic relationships between organs and tumors, facilitating the learning of structured features. Secondly, by replacing the conventional output layer with a lightweight class-specific segment head (CSH) integrated with language embeddings, our model achieves precise segmentation of 25 organs and 6 tumor types. The framework design allows for efficient extensibility, enabling the addition of new classes without catastrophic forgetting. To evaluate the effectiveness of our proposed framework, we conducted extensive experiments on a comprehensive dataset assembly. The training set comprised 3410 CT volumes from 14 different datasets, and the model was evaluated on 6162 external CT volumes from three additional datasets. Our framework achieved first place on six CT tasks in the Medical Segmentation Decathlon (MSD) public leaderboard and demonstrated state-of-the-art results on the Beyond The Cranial Vault (BTCV) dataset. Furthermore, the proposed Universal Model exhibited strong generalization to CT volumes from different centers, superior transfer learning capabilities across various downstream segmentation tasks, and outstanding continual learning capacity on two extension benchmarks. It also demonstrated computational efficiency, being six times faster than dataset-specific models.

2. Related work

2.1. Organ segmentation and tumor detection

Organ segmentation and tumor detection, which are essential components of radiotherapy, offer quantitative insights for diagnosing and categorizing diseases. Currently, deep learning techniques have gained considerable popularity in this domain due to their ability to extract data-driven features and facilitate end-to-end training. Among these techniques, U-Net (Ronneberger et al., 2015) and its variations (Zhou et al., 2019b; Liang et al., 2021; Oktay et al., 2018; Isensee et al., 2021) have emerged as prominent approaches, exhibiting promising outcomes. In addition to these convolutional neural networks (CNN)-based methods, generative adversarial network (GAN)-based approaches were also proposed to achieve high accuracy in segmentation tasks by employing adversarial losses (Cai et al., 2019; Gao et al., 2021; Mahmood et al., 2019). Nevertheless, training GAN networks is challenging and time-consuming since the generator must attain Nash equilibrium with the discriminator. Recently, transformer-based models (Zhou et al., 2021a; Hatamizadeh et al., 2022b; Tang et al., 2022; Chen et al., 2021a; He et al., 2023) use attention mechanism to learn the long-range relationship with stronger modeling capacities. Transformer-based approaches can capture long-range dependencies and achieve better performance than CNNs in many tasks.

These works often simplify multi-organ segmentation and tumor detection, focusing on specific organs and tumors (Chen et al., 2023a; Ronneberger et al., 2015; Zhou et al., 2019b; Isensee et al., 2021; Liang et al., 2021; Li et al., 2023), which have been verified to be effective for segmenting the organs and localizing the tumors. However, these works ignore the correlation between organs and tumors, specifically the prior anatomical information. Unlike these works, Universal Model advances a single framework that utilizes the large language model embedding to capture the semantic relationship between organs and tumors, addressing both tasks with extensibility. Moreover, we demonstrate our work on publicly available datasets, which is beneficial to reproducibility.

2.2. Large language vision model

With the widespread success of large models in the field of language processing (Brown et al., 2020; Devlin et al., 2018), there have been significant advancements in the application of language image models to visual recognition problems (Lüddecke and Ecker, 2022; Rao

et al., 2022; Wang et al., 2022a; Park et al., 2022; Xie et al., 2022a). Some language image models adopt BERT-like architecture (Yan and Pei, 2022; Conneau and Lample, 2019) or adopt contrastive learning paradigm (Radford et al., 2021; Wang et al., 2022b) to deal with vision and language data. While these work focus on pre-trained vision language models, another line of works focus on integrating the pre-trained language model into visual recognition problems. These models are capable of image captioning (Hu et al., 2022), visual question answering (Eslami et al., 2023), pathology image analysis (Huang et al., 2023), and so on. Qin et al. (2022) suggested that a language model could be used for detection tasks in the medical domain with carefully designed medical prompts. Grounded in these findings, we are among the first to introduce large language embedding to voxel-level semantic understanding medical tasks, i.e., segmentation, in which we underline the importance of the semantic relationship between anatomical structures.

2.3. Incremental learning

Incremental learning aims to acquire knowledge about new classes in multiple stages without experiencing catastrophic forgetting (Lewandowsky and Li, 1995), even when previous data is not accessible due to medical privacy regulations. This ability to extend a model to new classes without relying on previous data is crucial for organ segmentation and tumor detection. In the medical field, there are only a few existing methods for continual segmentation. For instance, Ozdemir et al. (2018) and Ozdemir and Goksel (2019) extended the distillation loss to address medical image segmentation. Another study by Liu et al. (2022b) introduced a memory module to store prototypical representations of different organ categories. However, these methods rely on a limited number of representative exemplars, which may not be practically available. In a recent work by Ji et al. (2023), the focus lies on network extension to tackle the forgetting problem. This is achieved by freezing the encoder and decoder and adding additional decoders when learning new classes. Although these approaches have successfully mitigated the forgetting issue, they incur significant memory costs in terms of network parameters. In contrast, Universal Model proposes a novel continual multi-organ and tumor segmentation method that overcomes the forgetting problem while imposing minimal memory and computation overhead.

2.4. Learning with integrated datasets

Due to financial constraints and clinical limitations, publicly available datasets for abdominal imaging primarily focus on a limited number of organs and tumors (Landman et al., 2017; Ma et al., 2021; Luo et al., 2021; Ji et al., 2022). For example, the AbdomenCT-1K dataset is designed for segmenting four organs (Ma et al., 2021), the WORD dataset for segmenting 16 organs (Luo et al., 2021), and the TotalSegmentor dataset for segmenting 104 anatomical structures (Wasserthal et al., 2022). To enhance representation capabilities, some studies have proposed simultaneous integration of multiple datasets (Zhou et al., 2019a; Fang and Yan, 2020; Zhang et al., 2021, 2022). However, integrating data can introduce challenges due to inconsistent label taxonomy, leading to the issue of partial labeling. Several approaches have been explored to leverage partial labels in organ segmentation and tumor detection (Liu et al., 2022a; Chen et al., 2021b; Bai and Xia, 2022; Zlocha et al., 2019; Yan et al., 2019; Liu et al., 2022c; Naga et al., 2022; Yan et al., 2020a; Mattikalli et al., 2022; Xie et al., 2023; Liu et al., 2024; Wu et al., 2022). For example, DoDNet (Zhang et al., 2021) and TransDoDNet (Xie et al., 2023) used one-hot embedding to generate task-specific output. COSST (Liu et al., 2024) utilized ground truth and pseudo-labels with self-training. TGNet (Wu et al., 2022) proposed task-guided attention modules and residual blocks to extract task-related features. Nonetheless, these studies have certain limitations. Firstly, the integrated datasets in these studies were relatively small-scale,[1] and the potential benefits of dataset integration were not convincingly demonstrated. Their performance was comparable to that of models trained on specific datasets, and lacked evaluation on official benchmarks. Secondly, the used one-hot embedding resulted in the disregard of the semantic relationship between organs and tumors.

2.5. Our previous work

We first presented CLIP-Driven Universal Model in our ICCV 2023 paper (Liu et al., 2023a) and our MICCAI 2023 paper (Zhang et al., 2023b), capable of segmenting 25 organs and seven types of tumors within CT volumes. Universal Model has since been quickly adopted by the research community, either as a strong baseline for comparison (Jiang et al., 2023; Chen et al., 2023b; Gao et al., 2023; Ye et al., 2023b; Ulrich et al., 2023), or as a source of inspiration for developing newer unified architectures (Liu et al., 2023b; Ye et al., 2023a; Zhang et al., 2023a; Zeng et al., 2023; Silva-Rodríguez et al., 2023); it has also been integrated into multiple open-source software, such as MONAI platform[2] at NVIDIA and Chimera/ChimeraX Group[3] at UCSF. To further strengthen Universal Model on our own, this paper presents **six extensions** to our previous work:

1. We design a new language-driven parameter generator with an independent multi-layer perceptron to avoid the entangling among various classes (Section 3.2.2).
2. We investigate the generalization capabilities of Universal Model on four external datasets without additional fine-tuning or adaptation (Fig. 4(a)).
3. We benchmark various partially labeled methods utilizing multiple datasets to evaluate the effectiveness of our model on the BTCV and MOTS datasets (Tables 4 and 5).
4. We further conduct case studies on challenging tumors, comparing the performance of AI with human experts (Fig. 5).
5. We perform inference in a clinical study setting to evaluate the model's capability in detecting small tumors that may have been overlooked during radiological evaluation (Section 4.5).
6. We analyze the model's transferability in segmenting sub-classes of pancreatic tumors, which represents a more challenging task requiring fine-grained discrimination (Table 9).

3. Methodology

3.1. Problem definition

Let us consider a set of partially labeled datasets $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N\}$ with corresponding organs/tumors label space $\{\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_N\}$, we learn a unified multi-organ segmentation model to segment tumors/organs in $\mathcal{L} = \bigcup_i^N \mathcal{L}_i$, where $\|\mathcal{L}\| \geq \|\mathcal{L}_i\|$ for all $i \in \{1, 2, \ldots, N\}$. When it comes a new dataset $\mathcal{D}_{N+1}$ with organs/tumors label space $\mathcal{L}_{N+1}$, $\mathcal{L}_{N+1} \setminus \mathcal{L} \neq \emptyset$, all previous training data are not accessible. The model is required to predict the accumulated label space $\mathcal{L} \bigcup \mathcal{L}_{N+1}$ for all seen datasets $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N, \mathcal{D}_{N+1}\}$.

[1] Zhou et al. (2019a) assembled 150 CT volumes from 4 datasets; Fang and Yan (2020) assembled 548 CT volumes from 4 datasets; Zhang et al. (2021) and Xie et al. (2023) assembled 1155 CT volumes from 7 datasets.

[2] MONAI: https://monai.io.

[3] UCSF ChimeraX: https://www.cgl.ucsf.edu/chimerax/.

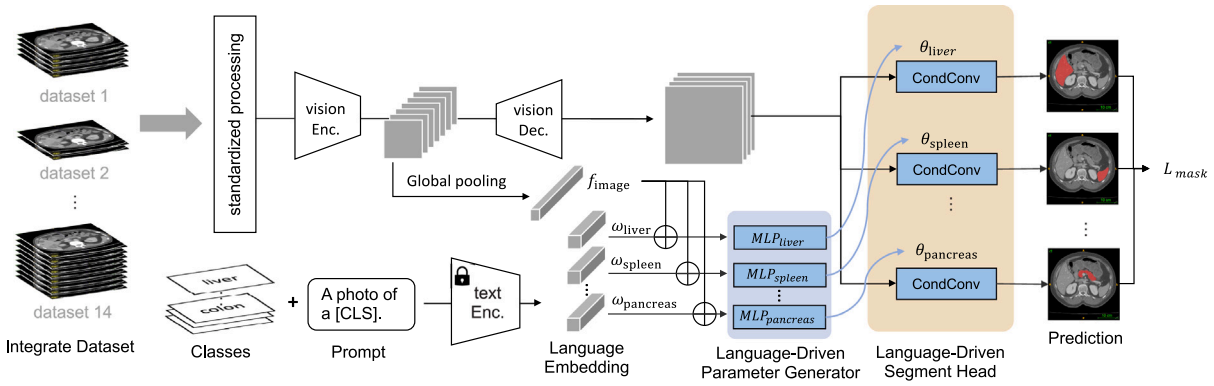


Fig. 1. *Overview.* We have developed the continual CLIP-Driven Universal Model from an assembly of 14 public datasets of 3410 CT volumes. In total, 25 organs and 6 types of tumors are partially labeled. To deal with partial labels, Universal Model consists of a language branch and a vision branch (Section 3.2). The official test set of MSD and BTCV are used to benchmark the performance of organ segmentation (Section 4.2) and tumor detection (Section 4.3). 3D-IRCADb, TotalSegmentator and a large-scale private dataset, consisting of 5038 CT volumes with 21 annotated organs, are used for independent, external validation of model generalizability (Section 4.5) and transferability (Section 4.6). The LPG module utilizes separate MLPs for each organ to overcome the entanglement issue present in the ICCV version (Liu et al., 2023a), which relied on a single MLP.

3.2. CLIP-driven universal model

As illustrate in Fig. 1, we present the framework of CLIP-Driven Universal Model for multi-organ segmentation and tumor detection with an integrated dataset, which consists of two main branches: the language branch and the vision branch. Given a 3D CT volume $\mathbf{X} \in D_n$, along with a partial label $\mathbf{Y}$, the language branch firstly generates the language embedding for each organ/tumor cls $\in \mathcal{L}_n$. To ensure that each parameter is aware of the anatomical relationship among classes, the language-driven parameter generator (LPG) consists of independent multi-layer perceptron (MLP) that take language embedding as input and generate a parameter θ for each class. On the other hand, the vision branch focuses on extracting the feature embedding $\boldsymbol{F}_E$ for the CT volume $\mathbf{X}$. Subsequently, the class-specific segment head (CSH) takes the generated parameter θ as parameter and $\boldsymbol{F}_D$ as input to predict the binary mask for each organ/tumor. During the continual learning process, the LPG module adopts another MLP to generate the new parameter and predict the new organ/tumor mask, minimizing the potential disruption caused by the introduction of new classes on the old classes.

3.2.1. Backbone network

For each 3D CT volume, isotropic spacing and uniformed intensity scale are utilized to reduce the domain gap among various datasets[4] The standardized and normalized CT volumes are subjected to the vision encoder and vision decoder. The choice of backbone architecture for this process is flexible and depends on the specific requirements of the task. Transformer-based architectures such as Swin UNETR or CNN-based architectures like U-Net are viable options, as long as they can effectively learn general representations for organ segmentation and tumor detection. The vision encoder extracts image features denoted as $\boldsymbol{F}_E$, then the vision decoder further refines the image features as $\boldsymbol{F}_D$. These features capture relevant information about the anatomical structures and potential abnormalities present in the volumes.

3.2.2. Language-driven parameter generator

To capture the semantic relations among organs/tumors, Universal Model utilizes the embedding of large language model, e.g., text encoder in CLIP (Radford et al., 2021)[5] and BioBERT (Yasunaga et al., 2022), to encode different classes since they can learn the semantic relationships present in abundant language data. To generate the language embedding $\boldsymbol{w}_{cls}$ for each class cls $\in \mathcal{L}_n$, the medical prompt is utilized, e.g., "a computerized tomography of a [CLS]", where [CLS] is the organ/tumor name of class cls. Subsequently, a global average pooling (GAP) operation is applied to the encoded image feature $\boldsymbol{F}_E$ to extract image-specific information, resulting in a global feature $\boldsymbol{f} = \text{GAP}(\boldsymbol{F}_E)$. The obtained global feature $\boldsymbol{f}$ and the language embedding are concatenated and fed into an independent multi-layer perceptron (MLP) to generate the corresponding parameter $\boldsymbol{\theta}_{cls}$,

$$\boldsymbol{\theta}_{cls} = \text{MLP}_{cls}(\boldsymbol{w}_{cls} \oplus \boldsymbol{f}), \tag{1}$$

where $\oplus$ denotes the concatenation. In this way, Universal Model exploits the semantic relationships among classes to address the label orthogonality problem, enhancing its ability to perform organ segmentation and tumor detection tasks.

3.2.3. Class-specific segment head

To address the label taxonomy inconsistency and enable efficient dynamic learning for different classes, we advance the Class-specific Segmentation Head for Universal Model framework. To achieve this, we employ conditional convolution, which involves three sequential convolutional layers with $1 \times 1 \times 1$ kernels and a non-linear function. The purpose of this approach is to generate a binary mask for each class. The first two layers consist of 8 channels, while the last layer has 1 channel. To initiate the process, we divide the parameter $\boldsymbol{\theta}_{cls}$ into three parts, namely $\boldsymbol{\theta}_{cls_1}$, $\boldsymbol{\theta}_{cls_2}$, and $\boldsymbol{\theta}_{cls_3}$. These parts are then assigned as parameters for the conditional convolution kernel. Next, we input the decoded feature $\boldsymbol{F}_D$ into the conditional convolution to obtain the binary mask,

$$\boldsymbol{P}_{cls} = \text{Sigmoid}(\phi(\phi(\boldsymbol{F}_D * \boldsymbol{\theta}_{cls_1}) * \boldsymbol{\theta}_{cls_2}) * \boldsymbol{\theta}_{cls_3}), \tag{2}$$

where $\phi(\cdot)$ represents a non-linear function, and the symbol * denotes the convolution operation. It is important to note that for each class cls, we generate the prediction using the "one vs. all" approach, employing the Sigmoid function instead of Softmax. This strategy allows for multi-label prediction, accommodating scenarios where a pixel may belong to multiple classes simultaneously (e.g., a tumor on an organ).

[4] The variations in acquisition parameters, reconstruction kernels, contrast enhancements, intensity variation and so on would lead to potential domain gap for different datasets (Orbes-Arteaga et al., 2019; Yan et al., 2020b; Guo et al., 2021a).

[5] CLIP (Contrastive Language–Image Pre-training), which has been trained on a vast dataset consisting of 400 million image–text pairs, including medical images and text (Chambon et al., 2022), exploits the semantic relationship between images and language.

Table 1
The information for an assembly of datasets. We have developed Universal Model from an assembly of 1–14 public datasets. The official test and validation sets of Medical Segmentation Decathlon (MSD) and Beyond the Cranial Vault (BTCV) are used to benchmark the performance of organ segmentation (Section 4.2) and tumor detection (Section 4.3). 3D-IRCADb (15), TotalSegmentator (16) and a large-scale private dataset (17), consisting of 5038 CT volumes with 21 annotated organs, are used for independent evaluation of model generalizability (Section 4.5) and transferability (Section 4.6).

Datasets (Year)	# Targets	# Volumes	Annotated organs or tumors
1. Pancreas-CT (2015)	1	82	Pancreas
2. LiTS (2019)	2	201	Liver, Liver Tumor*
3. KiTS (2020)	2	300	Kidney, Kidney Tumor*
4. AbdomenCT-1K (2021)	4	1000	Spleen, Kidney, Liver, Pancreas
5. CT-ORG (2020)	4	140	Lung, Liver, Kidneys and Bladder
6. CHAOS (2018)	4	40	Liver, Left Kidney, Right Kidney, Spl
7-11. MSD CT Tasks (2021)	9	947	Spl, Liver and Tumor*, Lung Tumor*, Colon Tumor*, Pan and Tumor*, Hepatic Vessel and Tumor*
12. BTCV (2015)	13	50	Spl, RKid, LKid, Gall, Eso, Liv, Sto, Aor, IVC, R&SVeins, Pan, RAG, LAG
13. AMOS22 (2022)	15	500	Spl, RKid, LKid, Gall, Eso, Liv, Sto, Aor, IVC, Pan, RAG, LAG, Duo, Bla, Pro/UTE
14. WORD (2021)	16	150	Spl, RKid, LKid, Gall, Eso, Liv, Sto, Pan, RAG, Duo, Col, Int, Rec, Bla, LFH, RFH
15. 3D-IRCADb (2010)	13	20	Liv, Liv Cyst, RLung, LLung, Venous, PVein, Aor, Spl, RKid, LKid, Gall, IVC
16. TotalSegmentator (2022)	104	1024	Clavicula, Humerus, Scapula, Rib 1-12, Vertebrae C1-7, Vertebrae T1-9, Vertebrae L1-5, Hip, Sacrum, Femur, Aorta, Pulmonary Artery, Right Ventricle, Right Atrium, Left Atrium, Left Ventricle, Myocardium, PVein, SVein, IVC, Iliac Artery, Iliac Vena, Brain, Trachea, Lung Upper Lobe, Lung Middle Lobe, Lung Lower Lobe, AG, Spl, Liv, Gall, Pan, Kid, Eso, Sto, Duo, Small Bowel, Colon, Bla, Autochthon, Iliopsoas, Gluteus Minimus, Gluteus Medius, Gluteus Maximus
17. JHH (2022)	21	5038	Aor, AG, CBD, Celiac AA, Colon, duo, Gall, IVC, Lkid, RKid, Liv, Pan, Pan Duct, SMA, Small bowel, Spl, Sto, Veins, Kid LtRV, Kid RtRV, CBD Stent, PDAC*, PanNET*, Pancreatic Cyst*

3.2.4. Optimization

The integrated dataset contains only partial annotations, resulting in some other organs being annotated as background. To mitigate this conflict, we employ the binary cross entropy (BCE) and binary dice loss functions for supervision. In order to generate the binary ground truth $\mathcal{M}_{cls}$ for each class within $\mathbf{L}_n$, we define the following formulation:

$$\mathbf{M}^i_{cls} = \begin{cases} 1, & \text{if } \mathbf{Y}^i = cls \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where i represents the voxel index. Subsequently, we calculate the BCE loss and dice loss summation for the classes contained in $\mathcal{L}_n$, and back-propagate the accurate supervision to update the entire framework. The formulation for this process is as follows:

$$\mathbf{L}_{mask} = \sum_{cls} Dice(\boldsymbol{P}_{cls}, \mathbf{M}_{cls}) + BCE(\boldsymbol{P}_{cls}, \mathbf{M}_{cls}), \text{for } cls \in \mathcal{L}_n. \tag{4}$$

The application of masked back-propagation effectively addresses the issue of label inconsistency in the context of the partial label problem.

3.3. Extended to novel classes

For the new CT volume $\{\mathbf{X}, \mathbf{Y}\}$ in $\mathcal{D}_{N+1}$, there are new classes $cls \in \mathcal{L}_{N+1} \backslash \mathcal{L}$. We add additional MLPs in LPG to learn the parameter for CSH and segment the new binary result for each new class. The feed forward pipeline follows Eqs. (1) and (2). The separate heads allow independent probability prediction for newly introduced and previously learned classes, therefore minimizing the impact of new classes on old ones during continual learning.

In the context of continual organ segmentation, a major challenge arises from the model's inability to access previous datasets, which often leads to forgetting of previously learned classes. To address this issue and preserve the existing knowledge, we propose a method of generating soft pseudo annotations for the old classes on newly-arrived data. Specifically, we leverage the output prediction from model denoted as $\hat{\mathbf{Y}}$, which includes the old classes $\mathcal{L}$. This prediction serves as the pseudo label for the current step's old classes. For the new classes, we continue to use the ground truth label. Formally, the binary ground truth $\mathbf{M}_{cls}$ for each class in continual learning can be expressed as:

$$\mathbf{M}_{cls} = \begin{cases} \mathbf{Y}_{cls} & \text{if } cls \in \mathcal{L}_{N+1} \setminus \mathcal{L} \\ \hat{\mathbf{Y}}_{cls} & \text{if } cls \in \mathcal{L}, \end{cases} \tag{5}$$

where $\mathbf{Y}_{cls}$ represents the ground truth label for class cls. By adopting this approach, our aim is to retain the previous knowledge and prevent the model from forgetting the information learned from the old classes while simultaneously learning the new classes.

4. Experiment I: Integrated learning

4.1. Experimental settings

4.1.1. Datasets

The assembly of datasets for this research comprises a total of 14 publicly available datasets, consisting of 3410 CT volumes, which were used for AI development—2100 for training and 1310 for validation. Additionally, 2 public datasets, 2 private datasets were employed for testing the developed model. The details of these datasets are summarized in Table 1. To preprocess the datasets, we initially map them to the standard index template, as shown in Table 2. For datasets such as KiTS, WORD, AbdomenCT-1K, and CT-ORG, where the left and right organs are not distinguished, we utilize a script to split the organs (Kidney, Adrenal Gland, and Lung) into left and right parts. Additionally, we account for the inclusion relation, such as considering organ tumors as part of the respective organs and the hepatic vessel being inside the liver. Since we represent each organ segmentation result as a binary mask, we can independently organize the segmentation ground truth for the overlapped organs in a binary mask format. This pipeline allows for efficient handling of the segmentation ground truth.

In addition to 16 public datasets, we also used two proprietary datasets collected from Johns Hopkins Hospital (JHH) and University of California, San Francisco (UCSF). JHH includes 5038 CT volumes with annotations for 21 organs. Each case in this dataset was scanned using contrast-enhanced CT in both venous and arterial phases, employing Siemens MDCT scanners. The JHH dataset is particularly employed to investigate the extensibility of new organ classes. The UCSF dataset comprises outpatient contrast-enhanced abdominal CT exams acquired at the portal venous phase from a single day (3/22/2024). These volumes were performed from multiple sites at UCSF, comprising a cohort of 11 patients with known liver metastases. These patients had varying numbers (ranging from 2 to 50) and sizes (ranging from 0.5 cm to 4.6 cm) of liver metastases.

4.1.2. Implementation details

Optimization. The training of Universal Model employs the AdamW optimizer, along with a warm-up cosine scheduler lasting for 50 epochs. For the segmentation experiments, a batch size of 6 per GPU is utilized, with a patch size of 96 × 96 × 96. The initial learning rate is set to $4e^{-4}$, with a momentum of 0.9 and a decay rate of $1e^{-5}$ on a multi-GPU setup consisting of 4 GPUs using Distributed Data Parallel (DDP). The implementation of the framework is based on MONAI 0.9.0. To

Table 2
The unification of the label taxonomy. The corresponding template for this taxonomy is presented, where ind. denotes the index of the class.

Class	Ind.	Class	Ind.	Class	Ind.
Spleen	1	R kidney	2	L kidney	3
Gall bladder	4	Esophagus	5	Liver	6
Stomach	7	Aorta	8	Postcava	9
Portal &spenic vein	10	Pancreas	11	R adrenal gland	12
L adrenal gland	13	Duodenum	14	Hepatic vessel	15
R lung	16	L lung	17	Colon	18
Intestine	19	Rectum	20	Bladder	21
Prostate/uterus	22	L head of femur	23	R head of femur	24
Celiac truck	25				
Kidney tumor	26	Liver tumor	27	Pancreatic tumor	28
Hepatic vessel tumor	29	Lung tumor	30	Colon tumor	31
Kidney cyst	32				

ensure the robustness of the results, a five-fold cross-validation strategy is employed. The best model is selected in each fold based on the evaluation of the validation metrics. The training process is conducted on eight NVIDIA RTX A5000 cards.

Data Augmentation. Data augmentation techniques are applied to enhance the diversity and generalization capability of the training dataset. In our implementation, we utilize the MONAI library (Cardoso et al., 2022) in Python. The uniform sampling strategy is utilized to ensure that volumes from each dataset have an equal probability of being selected. The preprocessing steps involve several operations to standardize the input CT volumes. Firstly, the orientation of the volumes is adjusted to conform to specified axcodes. Additionally, isotropic spacing is employed to reslice each volume, ensuring a consistent voxel size of $1.5\times1.5\times1.5$ mm^3. To normalize the intensity values, we truncate the range to $[-175, 250]$ and perform linear normalization to map the intensity values to the interval $[0, 1]$. As we focus on the relevant structures within the medical images, we perform cropping to extract the foreground objects with a foreground-to-background patch ratio of 2:1. Specifically, during the training phase, random fixed-sized regions of $96 \times 96 \times 96$ are cropped, with the center of the patch selected based on a predefined ratio. The selection process takes into account whether the center voxel corresponds to a foreground or background region. Also, we introduce random rotations of the input patches by 90 degrees and apply intensity shifts with a probability of 0.1 and 0.2, respectively. It is worth noting that mirroring augmentation is not utilized to avoid confusion between organ structures in the right and left parts of the images.

Network Structures. We use the pre-trained text encoder with masked self-attention transformer architecture from the CLIP framework as the text branch.[6] We can extract and store the text features to reduce overhead brought by the text encoder in the training and inference stage since the CLIP embedding only depends on the dictionary, which is fixed. For the vision branch, we employ Swin UNETR as our vision encoder. The Swin UNETR architecture consists of four attention stages, each composed of two transformer blocks, as well as five convolution stages utilizing a CNN-based structure. To reduce the resolution by a factor of 2 in the attention stage, a patch merging layer is utilized. Specifically, in Stage 1, a linear embedding layer and transformer blocks are employed to maintain the number of tokens at $\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2}$. Following this, a patch merging layer groups patches with a resolution of $2 \times 2 \times 2$ and concatenates them, resulting in a 4C-dimensional feature embedding. Subsequently, a linear layer is used to down-sample the resolution by reducing the dimension to 2C. This process continues in stages 2, 3, and 4, as outlined in Tang et al. (2022). The text-based controller, on the other hand, is implemented as a single convolutional layer. It takes as input the CLIP embedding and the global pooling feature from the last convolution stages in the vision encoder. This design choice enables efficient integration of text and vision information within the framework.

Inference. The input 3D image is divided into smaller overlapping windows or patches based on the (96, 96, 96)px region of interest and 0.5 overlap ratio. Each window is passed through the pre-trained model, and the model makes a prediction for that particular window. Then, the predictions from overlapping windows are aggregated using Gaussian weighted blending, which gives less weight to predictions on edges of windows, to form the final prediction for the entire input image. The obtained binary prediction is refined through three post-processing techniques. (1) Left-Right Split: For organs with bilateral symmetry, we fit a hyperplane through the vertebral and sagittal centers to split the left and right sides, subsequently remapping the side-related labels based on this hyperplane. (2) Non-largest Connected Component Suppression: For organs that occur only once, we identify three-dimensional connected components for the anatomy and remove non-largest connected components. (3) Anatomical Region Restriction: For certain categories with specific associated regions, we restrict the anatomy predictions to the corresponding body part. Since one pixel may belong to many classes at the same time, such as liver, liver tumor and Hepatic Vessel, it is optimal to store the prediction with binary mask for each classes.

When merging the binary masks into a single mask, we follow a specific order to merge the individual binary masks. Firstly, the organ masks are combined, followed by the incorporation of vessel masks, and finally, the tumor masks are integrated. Moreover, in instances where the predicted masks of two organs overlap, we adjudicate the final result by comparing the associated confidence scores for each prediction.

4.2. Organ segmentation benchmark

We secured the top #1 solution in two prominent medical segmentation challenges: Medical Segmentation Decathlon (MSD)[7] and Beyond The Cranial Vault (BTCV), surpassing the runners-up by a significant margin. Notably, our Universal Model demonstrates exceptional performance by providing solutions for six CT tasks, while predicting the results for four MRI tasks using open-sourced nnU-Net (Isensee et al., 2021).

4.2.1. Medical Segmentation Decathlon (MSD)

To assess the effectiveness of our approach, we conducted a comprehensive evaluation of both the official test set and a 5-fold cross-validation on the Medical Segmentation Decathlon (MSD) dataset. Detailed comparisons of our model's performance across various metrics are presented in Table 3, providing a comprehensive overview of its strengths. Specifically, for the liver, pancreas, hepatic vessel, and spleen organs, Universal Model achieved DSC scores of 95.42%, 82.84%, 67.15%, and 97.27%, respectively. Furthermore, Universal

[6] https://github.com/openai/CLIP.

[7] decathlon-10.grand-challenge.org/evaluation/challenge/leaderboard/.

Table 3
Leaderboard performance of CT tasks on MSD. The results are evaluated in the server on the MSD competition test dataset. The dice similarity coefficient (DSC) and normalized surface distance (NSD) metrics are obtained from the MSD public leaderboard. The results of MRI-related tasks were generated by nnU-Net (Isensee et al., 2021).

Method	Task03 liver						Task07 pancreas					
	DSC			NSD			DSC			NSD		
	Organ	Tumor	Avg.	Organ	Tumor	Avg.	Organ	Tumor	Avg.	Organ	Tumor	Avg.
Kim et al. (2019)	94.25	72.96	83.61	96.76	88.58	92.67	80.61	51.75	66.18	95.83	73.09	84.46
Trans VW (Haghighi et al., 2021)	95.18	76.90	86.04	97.86	92.03	94.95	81.42	51.08	66.25	96.07	70.13	83.10
C2FNAS (Yu et al., 2020)	94.98	72.89	83.94	98.38	89.15	93.77	80.76	54.41	67.59	96.16	75.58	85.87
Models Gen. (Zhou et al., 2021b)	95.72	77.50	86.61	98.48	91.92	95.20	81.36	50.36	65.86	96.16	70.02	83.09
nnU-Net (Isensee et al., 2021)	**95.75**	75.97	85.86	98.55	90.65	94.60	81.64	52.78	67.21	96.14	71.47	83.81
DiNTS (He et al., 2021)	95.35	74.62	84.99	**98.69**	91.02	94.86	81.02	55.35	68.19	96.26	75.90	86.08
Swin UNETR (Tang et al., 2022)	95.35	75.68	85.52	98.34	91.59	94.97	81.85	58.21	70.71	96.57	79.10	87.84
Universal Model	95.42	**79.35**	**87.39**	98.18	**93.42**	**95.80**	**82.84**	**62.33**	**72.59**	**96.65**	**82.86**	**89.76**

Method	Task08 hepatic vessel						Task06 lung		Task09 spleen		Task10 colon	
	DSC			NSD			DSC	NSD	DSC	NSD	DSC	NSD
	Organ	Tumor	Avg.	Organ	Tumor	Avg.	Tumor		Organ		Tumor	
Kim et al. (2019)	62.34	68.63	65.49	83.22	78.43	80.83	63.10	62.51	91.92	94.83	49.32	62.21
Trans VW (Haghighi et al., 2021)	65.80	71.44	68.62	84.01	80.15	82.08	74.54	76.22	97.35	99.87	51.47	60.53
C2FNAS (Yu et al., 2020)	64.30	71.00	67.65	83.78	80.66	82.22	70.44	72.22	96.28	97.66	58.90	72.56
Models Gen. (Zhou et al., 2021b)	65.80	71.44	68.62	84.01	80.15	82.08	74.54	76.22	97.35	99.87	51.47	60.53
nnU-Net (Isensee et al., 2021)	66.46	71.78	69.12	84.43	80.72	82.58	73.97	76.02	**97.43**	**99.89**	58.33	68.43
DiNTS (He et al., 2021)	64.50	71.76	68.13	83.98	81.03	82.51	74.75	77.02	96.98	99.83	59.21	70.34
Swin UNETR (Tang et al., 2022)	65.69	72.20	68.95	84.83	81.62	83.23	76.60	77.40	96.99	99.84	59.45	70.89
Universal Model	**67.15**	**75.86**	**71.51**	**84.84**	**85.23**	**85.04**	**80.01**	**81.25**	97.27	99.87	**63.14**	**75.15**

Model demonstrated superior performance in the segmentation of liver tumors, pancreas tumors, hepatic vessel tumors, lung tumors, and colon tumors, with DSC scores of 79.35%, 62.33%, 75.86%, 80.01%, and 63.14%, respectively.

4.2.2. Beyond The Cranial Vault (BTCV)

Table 4 presents a comparison between Universal Model and other methods under dataset-specific training and dataset-agnostic training on the Biomedical Translational Clinical Vision (BTCV) benchmark. Universal Model surpassed the performance of dataset-specific training methods by a minimum of 3.5% across various evaluation metrics. Furthermore, it outperformed dataset-agnostic training approaches, achieving an average Dice score superiority of 0.38%. Overall, Universal Model consistently exhibits strong performance on both the MSD and BTCV benchmarks outperforming the performance of alternative methods. These results affirm the robustness and effectiveness of Universal Model for organ segmentation across multiple tasks.

4.2.3. Multi-Organ and Tumor Segmentation (MOTS)

To facilitate an equitable comparison of Universal Model with other partial label learning methods, we utilized the Multi-Organ and Tumor Segmentation (MOTS) dataset (Zhang et al., 2021) as a benchmark. The MOTS dataset comprises seven partially labeled sub-datasets, encompassing seven organ and tumor segmentation tasks. It consists of 1155 CT volumes, with 920 volumes designated for the training set and 235 for the test set. We conducted separate training procedures for Universal Model on the MOTS training set and the 14 integrated datasets in this work (excluding the 235 CT volumes present in the MOTS test set). The results presented in Table 5 demonstrate the effectiveness of our proposed framework, achieving the second-best performance compared to other state-of-the-art methods. Moreover, when trained on a larger integrated dataset, Universal Model exhibited improved performance for most organs and tumors, validating the efficacy of training models on extensive integrated datasets.

4.3. Tumor detection on five datasets

Usually, tumor detection task solely relies on DSC scores, which may not accurately reflect the performance. DSC scores are typically calculated only on abnormal CT volumes that contain tumors, which can lead to an overestimation of performance (Isensee et al., 2021; Xia et al., 2022). The presence of normal CT volumes without tumors may result in numerous false positives generated by the AI system (Shen et al., 2021). To address this limitation, we evaluate the Sensitivity and Specificity at the patient level for the detection of three types of tumors. The harmonic mean of sensitivity and specificity is reported to provide a balanced assessment of the model's abilities. To obtain normal CT volumes without tumors for evaluation, we utilize the CHAOS and Pancreas-CT datasets, which provide pathological verification of the absence of tumors (Valindria et al., 2018; Roth et al., 2015). Table 6 show that Universal Model achieves a harmonic mean of 91.84% for liver tumors, 93.31% for kidney tumors, and 92.59% for pancreatic tumors, indicate that the proposed Universal Model can accurately identify tumor cases while reducing false positives, highlighting its clinical significance.

Moreover, we visualize several CT volumes from small to large size tumors in Fig. 2 to demonstrate the superior capacity of Universal Model. The visualization showcases the proficiency of Universal Model in detecting both small and large size pancreatic tumors, while other methods may either ignore them or only detect a portion of the tumor. In addition, row presents a CT volume from another hospital, where Universal Model demonstrates excellent generalization in pancreas organ segmentation and avoids generating false positives for tumors. This emphasizes the effectiveness of Universal Model in diverse scenarios. Compared with dataset-specific models, the smaller number of false positives predicted by our Universal Model underlines the necessity of assembling diverse datasets, benefiting from not only sufficient positive examples for training but also a larger number of negative examples as a control.

Overall, Universal Model demonstrates superior performance in terms of specificity and harmonic mean compared to the evaluated methods, indicating its ability to achieve a balanced performance in tumor detection across different organs.

4.4. Ablation study

4.4.1. Language encoding

In order to assess the effectiveness of language encoding, we conducted an experiment using different language models and varying prompt templates as shown in Table 7. The results indicate that the CLIP text encoder equipped with the prompt template "A computerized tomography of a [CLS]", achieves the highest performance with an

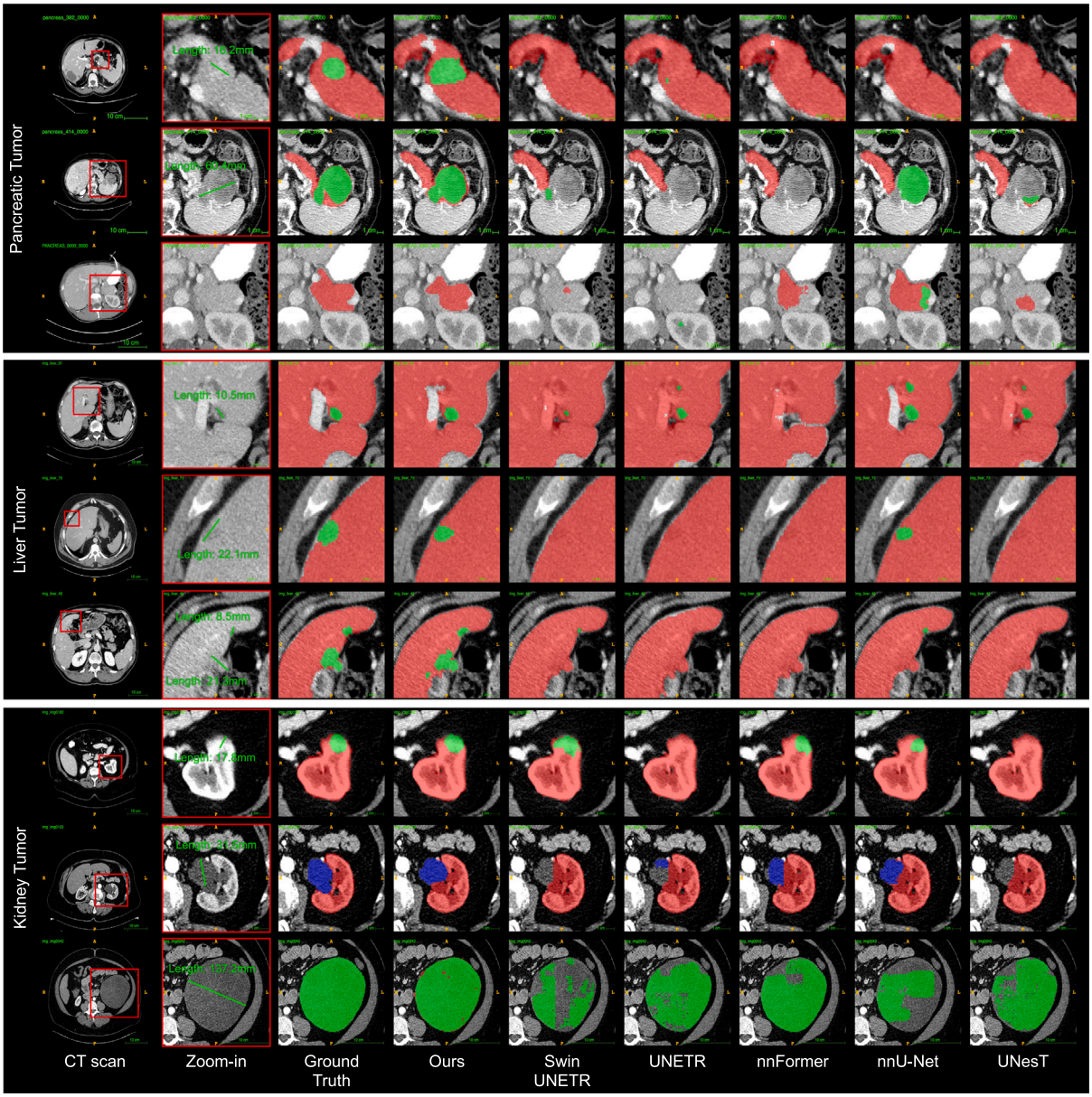


Fig. 2. *Qualitative results of multi-tumor detection and segmentation.* We review the detection/segmentation results of each tumor type from smaller to larger sizes. Especially, Universal Model generalizes well in organ segmentation and does not generate many false positives of tumors when it comes to a CT volume without tumors from other hospitals (Row 3).

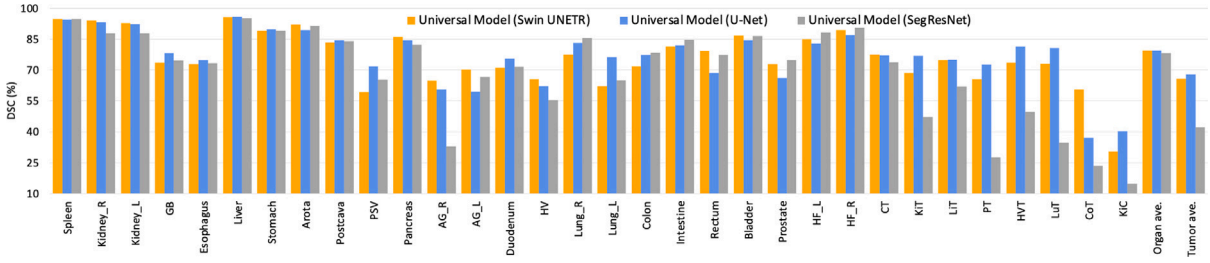
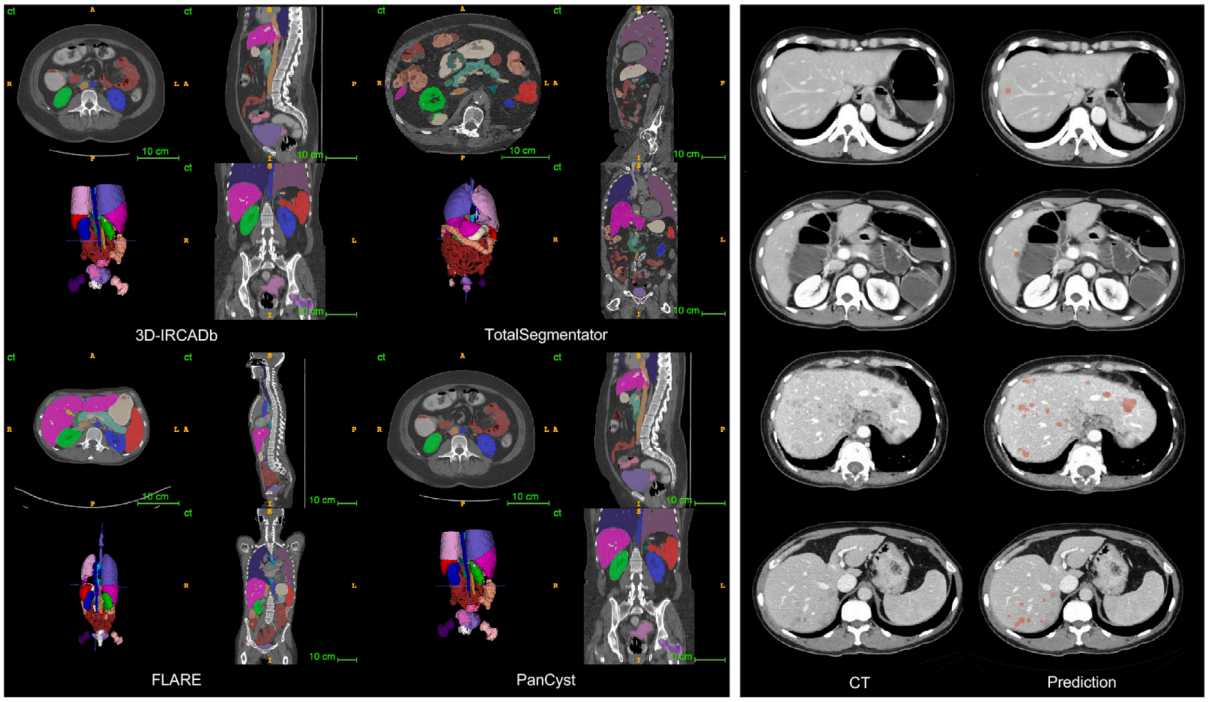


Fig. 3. *Ablation study on different segmentation backbones.* Universal Model can be expanded to Transformer-based (Swin UNETR) and CNN-based (U-Net, SegResNet-Tiny) backbone. These backbones achieve comparable results. The numbers of parameters of Swin UNETR, U-Net, and SegResNet-Tiny are 62.19M, 19.08M, and 4.7M, respectively. The order of classes is the same as Table 2.

average Dice score of 76.11%. It is worth noting that other language models and prompt templates also outperformed the one-hot encoding technique, which provides compelling evidence regarding the efficacy of language encoding in our study.

4.4.2. *Different backbones*

Universal Model framework can be applied flexibly to other backbones. We conduct experiments in two CNN-based backbones, namely U-Net (Ronneberger et al., 2015) and SegResNet (Myronenko, 2019).



(a) External validation on publicly available datasets (b) External validation on proprietary UCSF datasets

Fig. 4. *(a) Pseudo-label visualization.* 25 organs and 6 tumors in four unseen datasets are visualized. *(b) External validation for liver tumor detection.* In Cases 1 and 2, Universal Model successfully identified small new liver tumors, that have been overlooked during radiological evaluation. In Cases 3 and 4, where multiple liver tumors were present, Universal Model detected them, resulting in improved diagnostic efficiency.

Table 4
Five-fold cross-validation results on BTCV. For a fair comparison, we did not use model ensemble during the evaluation. The dataset-specific training means train only on BTCV, while dataset-agnostic training means training on multiple datasets. In dataset-agnostic training part, results of Swin UNETR and Multi-Talent are original from Ulrich et al. (2023). We re-implement TansUNet and UniSeg with same setting as Universal Model. Universal Model achieves the overall best performance. The performance is statistically significant at the P = 0.05 level, with highlighting in a light red box.

Methods	Spl	RKid	LKid	Gall	Eso	Liv	Sto	Aor	IVC	Veins	Pan	AG	Avg.
Dataset-specific training													
TransUNet (Chen et al., 2021a)	94.10	90.22	88.84	65.49	73.19	93.24	80.85	87.47	80.48	71.47	74.26	64.76	79.16
CoTr (Xie et al., 2021)	95.51	88.03	89.19	68.49	75.83	95.93	81.84	89.01	82.32	73.39	75.12	65.78	80.48
TransBTS (Isensee et al., 2021)	94.59	89.23	90.47	68.50	75.59	96.14	83.72	88.85	82.28	74.25	75.12	66.74	80.94
nnFormer (Zhou et al., 2021a)	94.51	88.49	93.39	65.51	74.49	96.10	83.83	88.91	80.58	75.94	77.71	68.19	81.22
UNETR (Hatamizadeh et al., 2022b)	94.91	92.10	93.12	76.98	74.01	96.17	79.98	89.74	81.20	75.05	80.12	62.60	81.43
nnU-Net (Isensee et al., 2021)	95.92	88.28	92.62	66.58	75.71	96.49	86.05	88.33	82.72	**78.31**	79.17	67.99	82.01
Swin UNETR (Tang et al., 2022)	95.44	93.38	93.40	77.12	74.14	96.39	80.12	90.02	82.93	75.08	81.02	64.98	82.06
Dataset-agnostic training													
TransUNet (Chen et al., 2021a)	94.90	90.89	90.45	70.82	80.31	95.01	90.11	91.30	84.21	74.53	83.21	71.27	84.75
Swin UNETR (Tang et al., 2022)	91.04	86.93	86.83	67.73	77.87	95.14	86.77	89.92	84.34	73.88	82.27	65.34	82.33
Multi-Talent (Ulrich et al., 2023)	93.61	90.89	90.47	72.12	**79.28**	96.23	**92.83**	91.60	**87.31**	77.58	**84.92**	**72.16**	85.75
UniSeg (Ye et al., 2023b)	**96.99**	94.11	92.21	68.13	78.15	96.43	85.93	89.78	86.69	76.12	83.39	72.39	85.02
Universal Model	95.82	**94.28**	**94.11**	**79.52**	76.55	**97.05**	92.59	**91.63**	86.00	77.54	83.17	70.52	**86.13**

Over 25 different organs, the framework achieved an average Dice Similarity Coefficient (DSC) score of 79.65% with U-Net and 78.30% with SegResNet. Furthermore, for 6 different tumor types, the framework achieved an average DSC score of 67.95% with U-Net and 42.20% with SegResNet. These scores are comparable to the performance of the Swin UNETR backbone with the average dice score of 79.56% and 65.79% for organs and tumors as shown in Fig. 3. The results highlight the versatility and effectiveness of the Universal Model as it demonstrates its ability to be expanded to different backbone architectures. The comparable performance achieved by these backbones further supports the flexibility and adaptability of Universal Model in various medical imaging tasks. In the future, more advanced backbones will be added over time in our GitHub repository.

4.5. Generalization for external datasets

The generalizability of medical AI models is a crucial expectation, as they should perform well on new data from various hospitals, rather than being tailored to a single dataset (Mongan et al., 2020; Norgeot et al., 2020; Guo et al., 2021b). In comparison to dataset-specific models, Universal Model was trained on a significantly larger and more diverse collection of CT volumes. This broader training enables Universal Model to demonstrate superior generalizability, as

Table 5
Benchmark of different partially labeled methods on MOTS dataset. Universal Model* denotes the model trained exclusively on the MOTS training set, while Universal Model refers to our proposed model trained on the 14 integrated datasets in this work, excluding the 235 CT volumes present in the MOTS test set. The average performance is statistically significant at the P = 0.05 level, with highlighting in a light red box.

Methods	Task 1: Liver		Task 2: Kidney		Task 3: Hepatic		Task 4: Pancreas		Task 5: Colon	Task 6: Lung	Task 7: Spleen	Average
	Organ	Tumor	Organ	Tumor	Organ	Tumor	Organ	Tumor	Tumor	Tumor	Organ	Dice
Multi-Nets	96.54	63.9	96.52	78.07	62.85	71.56	83.18	56.17	42.39	61.68	94.37	73.38
Cond-Input (Chen et al., 2017)	96.67	65.61	96.97	83.17	65.14	74.98	84.24	63.32	46.03	69.67	95.15	76.45
Multi-Head (Chen et al., 2019b)	96.77	64.33	96.84	82.39	65.21	74.62	84.59	64.11	46.65	68.63	95.47	76.33
Cond-Dec (Dmitriev and Kaufman, 2019)	96.23	65.25	96.43	83.43	65.38	72.26	83.86	62.97	50.67	63.77	94.33	75.87
TAL (Fang and Yan, 2020)	96.21	64.1	96.01	78.87	64.64	73.87	83.39	60.93	45.14	67.59	94.72	75.04
DoDNet (Zhang et al., 2021)	96.86	65.99	97.31	83.45	65.80	77.07	85.39	60.22	47.66	72.65	93.94	76.94
TGNet (Wu et al., 2022)	96.83	63.84	96.17	81.05	62.59	74.65	83.30	60.71	47.37	61.69	94.18	74.76
CCQ (Liu et al., 2023c)	96.71	64.32	96.68	79.82	62.50	**76.94**	83.18	60.54	54.77	70.51	94.57	76.41
TransDoDNet (Xie et al., 2023)	**97.01**	66.27	**97.03**	83.57	65.43	76.78	84.88	64.63	58.64	**70.82**	95.82	78.26
Universal Model*	96.93	67.14	96.09	82.34	64.39	76.31	84.42	64.68	59.49	70.31	95.31	77.95
Universal Model	96.81	**70.18**	96.96	**87.87**	**66.15**	76.45	**85.16**	**67.21**	**63.92**	69.63	**96.84**	**79.74**

Table 6
Quantitative results of multi-tumor detection. The tumor detection performance analysis is based on CT volumes from the LiTS, KiTS, and MSD Pancreas datasets, containing tumors in the liver, kidney, and pancreas, respectively (Bilic et al., 2019; Heller et al., 2019; Antonelli et al., 2021). These volumes are utilized to compute the sensitivity of tumor detection. To assess specificity in an alternative manner, the CHAOS and Pancreas-CT datasets are employed (Valindria et al., 2018; Roth et al., 2015). Importantly, it has been confirmed that the CHAOS dataset is devoid of liver or kidney tumors, while the Pancreas-CT dataset does not contain pancreatic tumors within the CT volumes. Universal Model achieves a high harmonic mean, which signifies the capability to accurately identify tumor cases while minimizing false positives.

Methods	Liver tumor			Kidney tumor			Pancreatic tumor		
	Sen.	Spec.	Harm.	Sen.	Spec.	Harm.	Sen.	Spec.	Harm.
nnU-Net (Isensee et al., 2021)	**94.44**	75.00	83.60	96.88	85.00	90.55	95.18	88.75	91.85
UNet++ (Zhou et al., 2019b)	**94.44**	80.00	86.62	N/A	N/A	N/A	N/A	N/A	N/A
UNETR (Hatamizadeh et al., 2022b)	86.11	**95.00**	90.34	93.75	**95.00**	**94.37**	90.36	81.25	85.56
Swin UNETR (Tang et al., 2022)	91.67	85.00	88.21	**97.91**	70.00	81.63	**97.59**	87.50	92.26
Universal Model	88.89	**95.00**	**91.84**	91.67	**95.00**	93.31	93.98	**91.25**	**92.59**

Table 7
Ablation study on language encoding. We compare the language embedding among one-hot, BioBERT (Yasunaga et al., 2022) and CLIP embedding in the test set of integrated data. The BioBERT method utilizes the prompt *'A computerized tomography of a [CLS]'*. On the other hand, CLIP v1, v2, and v3 employ prompts such as *'A photo of a [CLS]'*, *'There is [CLS] in this computerized tomography'*, and *'A computerized tomography of a [CLS]'*, respectively.

Encoding	Spleen	KidneyR	KidneyL	Gallblad.	Esophagus	Liver	Stomach	Aorta	Postcava	PSV	Pancreas
One-hot	91.92	91.98	92.14	71.75	70.28	95.10	80.52	83.57	82.71	67.81	74.06
BioBERT	94.65	93.26	**92.98**	75.14	72.32	95.09	87.68	91.05	**83.91**	67.83	80.51
CLIP V1	92.35	91.83	91.89	72.45	71.38	90.23	73.07	86.77	78.17	74.00	74.91
CLIP V2	93.05	92.14	91.42	**75.88**	**75.56**	94.75	75.79	91.15	80.64	**78.90**	78.94
CLIP V3	**94.69**	**94.09**	92.77	73.45	72.87	**95.71**	**89.19**	**92.19**	83.44	59.20	**86.09**
Embedding	AdrenalR	AdrenalL	Duodenum	Hepatic	LungR	LungL	Colon	Intestine	Rectum	Bladder	Prostate
One-hot	64.52	66.96	55.66	71.03	**79.63**	66.75	69.22	78.05	69.87	76.74	66.15
BioBERT	65.94	68.72	**68.61**	59.14	75.40	69.09	71.24	**81.78**	65.58	74.51	69.51
CLIP V1	72.07	72.42	62.42	**74.53**	79.32	76.52	70.32	75.65	63.11	75.06	66.47
CLIP V2	**79.98**	**79.73**	66.01	68.65	75.87	**82.98**	**74.88**	70.82	64.64	70.06	68.8
CLIP V3	64.75	70.18	71.11	65.43	77.48	62.11	71.77	81.47	**79.42**	**86.71**	**72.96**
Embedding	FemurL	FemurR	Celiac	KiT	LiT	PT	HVT	LuT	CoT	KiC	Ave
One-hot	70.27	60.23	78.92	63.84	68.02	55.48	52.31	53.87	48.39	**35.81**	70.42
BioBERT	74.39	79.07	80.69	57.41	63.44	39.70	57.88	58.57	54.19	20.33	71.55
CLIP V1	74.61	72.53	79.28	56.62	76.24	61.05	56.49	73.60	55.03	32.87	73.49
CLIP V2	69.98	75.73	**84.04**	67.04	**82.09**	**77.75**	67.45	**75.38**	55.55	35.79	75.66
CLIP V3	**84.94**	**89.45**	77.55	**68.72**	74.87	65.46	**73.53**	73.12	**60.66**	30.44	**76.11**

it can be directly tested on external data without the need for adaptation or fine-tuning. We conducted an assessment using both a public dataset, 3D-IRCADb, and a private dataset, JHH. These datasets were not included in the training phase and can be considered as external validation. The results, presented in Table 8, demonstrate that Universal Model outperforms previous methods on both 3D-IRCADb and JHH, with a substantial improvement in DSC similarity coefficient (DSC) of 5% and 4%, respectively.

Additionally, we conducted a pseudo-label visualization of 25 organs in four unseen datasets, including 3D-IRCADb, TotalSegmentator, FLARE, and PanCyst. This visualization, depicted in Fig. 4(a), provides further evidence of the performance on these unseen datasets achieved by Universal Model. For tumor detection, we opted for the private dataset from UCSF for external validation. Universal Model served as an adjunctive tool in radiologist diagnostic practices. Upon evaluation by a radiologist, Universal Model significantly reduced the false negative rate of liver tumors and improved diagnostic efficiency, particularly in the following clinical scenarios: 1. Detection of new small liver lesions, which are often overlooked but critical for initial tumor staging. 2. Assessment of changes in multiple tumors, which is a

Table 8
Generalizability. Results on external datasets. We evaluate Universal Model and eight models on data from two external sources without additional fine-tuning or domain adaptation. mDSC* is the average DSC score of the first seven organs. Compared with dataset-specific models, our Universal Model performs more robustly to CT volumes taken from a variety of scanners, protocols, and institutes. The average performance is statistically significant at the P = 0.05 level, with highlighting in a light red box.

3D-IRCADb	Spleen	KidneyR	KidneyL	Gallblad.	Liver	Stomach	Pancreas	LungR	LungL	mDSC*	mDSC
SegResNet (Siddiquee and Myronenko, 2021)	94.08	80.01	91.60	69.59	95.62	**89.53**	79.19	N/A	N/A	85.66	N/A
nnFormer (Zhou et al., 2021a)	93.75	88.20	90.11	62.22	94.93	87.93	78.90	N/A	N/A	85.14	N/A
UNesT (Yu et al., 2022)	94.02	84.90	**94.95**	68.58	95.10	89.28	79.94	N/A	N/A	86.68	N/A
TransBTS (Wang et al., 2021)	91.33	76.22	88.87	62.50	94.42	85.87	63.90	N/A	N/A	80.44	N/A
TransUNet (Chen et al., 2021a)	94.09	82.07	89.92	63.07	95.55	89.12	79.53	N/A	N/A	84.76	N/A
UNETR (Hatamizadeh et al., 2022b)	92.23	91.28	94.19	56.20	94.25	86.73	72.56	91.56	93.31	83.92	85.81
Swin UNETR (Tang et al., 2022)	93.51	66.34	90.63	61.05	94.73	87.37	73.77	93.72	92.17	81.05	83.69
Universal Model	**95.76**	**94.99**	94.42	**88.79**	**97.03**	89.36	**80.99**	**97.71**	**96.72**	**91.62**	**92.86**
JHH	Spleen	KidneyR	KidneyL	Gallblad.	Liver	Stomach	Pancreas	Aorta	Postcava	Vein	mDSC
SegResNet (Siddiquee and Myronenko, 2021)	93.11	89.92	87.84	74.62	95.37	87.90	76.33	84.05	79.36	57.13	82.56
nnFormer (Zhou et al., 2021a)	86.71	87.03	84.28	63.37	91.64	73.18	71.88	84.73	78.61	55.31	77.67
UNesT (Yu et al., 2022)	93.82	90.42	89.04	76.40	95.30	89.65	78.97	84.36	79.61	59.70	83.73
TransBTS (Wang et al., 2021)	85.47	81.58	82.00	60.58	92.50	72.29	63.25	83.47	75.07	55.38	75.16
TransUNet (Chen et al., 2021a)	94.63	89.86	89.61	77.28	95.85	88.95	79.98	85.06	**81.02**	**59.76**	84.20
UNETR (Hatamizadeh et al., 2022b)	91.89	89.07	87.60	66.97	91.48	83.18	70.56	82.92	75.20	57.53	79.64
Swin UNETR (Tang et al., 2022)	92.23	84.34	82.95	74.06	94.91	82.28	71.17	**85.50**	79.18	55.11	80.17
Universal Model	**93.94**	**91.53**	**90.21**	**84.15**	**96.25**	**92.51**	**82.72**	77.35	79.64	57.10	**84.54**

Table 9
Transferability: Fine-tuning performance. Comparing with MedicalNet (Chen et al., 2019b), Models Gen. (Zhou et al., 2019c), Swin UNETR (Tang et al., 2022), UniMiSS (Xie et al., 2022b), fine-tuning Universal Model significantly outperforms learning from scratch on several downstream tasks (i.e., , vertebrae, cardiac, muscles, organs and sub-classes of pancreatic tumor segmentation). Moreover, Universal Model, trained by image segmentation as proxy task, can extract better visual representation—more related to segmentation tasks—than other pre-trained models developed in the medical domain. The average performance is statistically significant at the P = 0.05 level, with highlighting in a light red box.

Method	Total_vert.	Total_card.	Total_mus.	Total_org.	Total_ave	JHH_card.	JHH_org.	JHH_ave	PDAC	Cyst	PanNet	Pancre_ave
Scratch	81.06	84.47	88.83	86.42	85.20	71.63	89.08	80.36	53.1	41.2	34.6	42.97
MedicalNet	82.28	87.40	91.36	86.90	86.99	58.07	77.68	67.88	N/A	N/A	N/A	N/A
ModelsGen.	85.12	86.51	89.96	85.78	86.84	**74.25**	88.64	81.45	N/A	N/A	N/A	N/A
Swin UNETR	86.23	87.91	92.39	88.56	88.77	67.85	87.21	77.53	53.4	41.6	35.4	43.47
UniMiSS	85.12	88.96	92.86	88.51	88.86	69.33	82.53	75.93	N/A	N/A	N/A	N/A
Ours	**86.49**	**89.57**	**94.43**	**88.95**	**89.86**	72.06	**89.37**	**80.72**	**53.6**	**49.2**	**45.7**	**49.5**

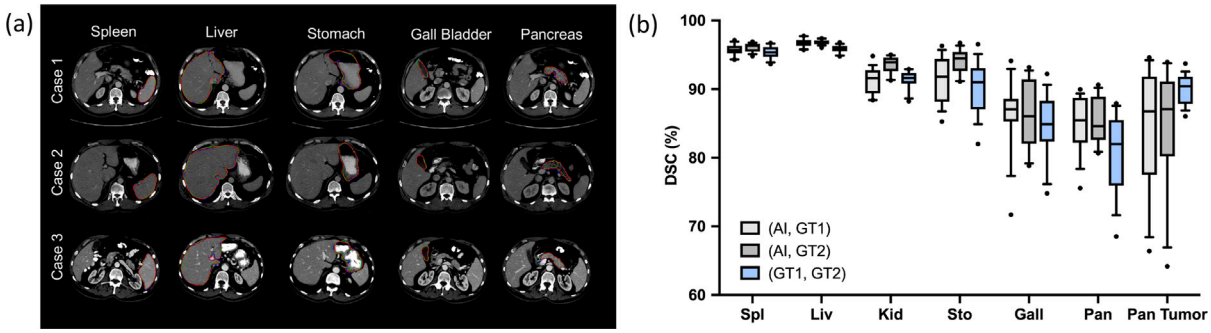


Fig. 5. *(a) Contour line comparison among pseudo labels and two human experts.* The red line represents the annotation from Doctor 1; green line indicates the annotation from Doctor 2; blue line shows the results generated by Universal Model. Examples of CT volumes annotated by our pseudo labels and two human experts with contour line comparison. The prediction results of these organs generated by the medical model are comparable with human experts. *(b) Intra-observer variability.* We obtain similar performance between pseudo labels generated by Universal Model (AI) and annotations performed by two human experts (Dr1,2) on 6 organs. Spleen (Spl), liver (Liv), kidneys (Kid), stomach (Sto), gallbladder (Gall), and pancreas (Pan) can be annotated by AI with a similar intra-observer variability to humans.

time-consuming and tedious task but essential for tumor prognosis. This visualization, depicted in Fig. 4(b), illustrates the preliminary results of liver tumor detection by Universal Model.

4.6. Fine-tuning results for other tasks

Universal Model possesses another significant characteristic, as it serves as a robust pre-training model for other tasks (Tang et al., 2024). By undergoing pre-training using an assembled dataset and subsequent fine-tuning on other datasets, Universal Model achieves the highest DSC score compared to other pre-training methods. Specifically, in the TotalSegmentator dataset, Universal Model achieves DSC scores of 86.49%, 89.57%, 94.43%, and 88.95% for four tasks, as presented in Table 9. This remarkable performance highlights the potential of Universal Model in enhancing the generalization of medical imaging models by effectively capturing fine-grained information for segmentation purposes. In addition, we investigated the cross-task ability of our model by transferring to the fine-grained segmentation of pancreatic tumors. We employed our proprietary dataset, which comprises 3577 annotated pancreatic tumors, including detailed sub-types: 1704 PDAC, 945 cysts, and 928 PanNets. The results shown in Table 9 also demonstrate that our model can transfer better to novel classes than self-supervised models, even in the more challenging sub-class tumor segmentation task.

4.7. Comparing AI predictions with human annotations

To compare the performance of AI model and human experts, we randomly select 17 CT volumes in BTCV. Two independent groups of radiologists from different institutes were involved in annotating six organs. The quality of the pseudo labels predicted by Universal Model and the manual annotations performed by human experts were assessed. Fig. 5(a) illustrates that the organ boundaries from both human experts and Universal Model exhibit a high degree of alignment. Additionally, it is worth noting that manual annotations performed by human experts exhibit inter-observer variance (Ji et al., 2021), particularly in segmentation tasks, which is attributed to the inherent blurriness and ambiguity of some organ boundaries. We further calculate the mutual DSC scores (DSC) for six organs and one tumor between the AI prediction and the two human annotations (GT1 and GT2), denoted as AI↔GT1, AI↔GT2, and GT1↔GT2, in Fig. 5(b). An intriguing observation was that the DSC for six organs between the AI model and human expert was slightly higher than the DSC among the human experts themselves. Notably, the AI model also achieved a smaller variance. However, the variance of the DSC for tumor segmentation between the AI model and human experts was substantially larger than that observed between the human experts, indicating that the model's tumor prediction capability still lags behind human expert proficiency. These findings underscore the model's proficiency in segmenting specific anatomical structures while concurrently underscoring the need for further refinements from human doctors to enhance its tumor segmentation during daily clinical application.

Furthermore, given that the six organs can be segmented by the AI model with a similar variance to human experts, it is encouraged for the research community to focus on creating annotations for more challenging organs and tumors.

5. Experiment II: Continual learning

5.1. Experimental settings

5.1.1. Datasets

In this study, we conducted two benchmark tests to evaluate the extensibility of the Continual Universal Model.

Tumor Extension Benchmark. The first benchmark setting was designed to simulate the tumor extension scenario encountered in clinical routine, and it utilized the BTCV dataset and LiTS dataset. The BTCV dataset consisted of 47 abdominal CT images that provided delineation for 13 different organs. Specifically, we utilized 13 classes from the BTCV dataset, namely spleen, right kidney, left kidney, gall bladder, esophagus, liver, stomach, aorta, inferior vena cava, portal vein and splenic vein, pancreas, right adrenal gland, and left adrenal gland, during the step 1 learning. For the step 2 learning, we employed the LiTS dataset, which comprised 130 contrast-enhanced abdominal CT volumes specifically for liver and liver tumor segmentation. In this stage, we focused on learning the liver tumor class from the LiTS dataset.

Body Extension Benchmark. The second benchmark setting aimed to simulate extension scenarios involving different body regions, and it utilized the in-house JHH dataset. The JHH dataset contained annotated data for multiple classes, which we categorized into three groups: abdominal organs, gastrointestinal tract, and cardiovascular system. In the abdominal organs category, we selected seven classes for the step 1 learning, including spleen, right kidney, left kidney, gall bladder, liver, postcava, and pancreas. In the gastrointestinal tract category, three classes (stomach, colon, and intestine) were learned during the step 2 learning. Finally, for the step 3 learning, we focused on learning four classes from the cardiovascular system, namely aorta, portal vein and splenic vein, celiac truck, and superior mesenteric artery. This categorization was performed in accordance with the TotalSegmentator (Wasserthal et al., 2022).

5.1.2. Baselines

For a fair comparison, all the compared methods use the same Swin UNETR (Hatamizadeh et al., 2022a) as the backbone network, the state-of-the-art model in many medical image segmentation tasks. We compare with three popular continual learning baseline methods that apply knowledge distillation, including LwF (Li and Hoiem, 2017), ILT (Michieli and Zanuttigh, 2019) and PLOP (Douillard et al., 2021).

5.1.3. Implementation details

The data augmentation, network structures, and evaluation metrics employed in this experiment are consistent with Section 4.

Optimization. We train the proposed network architecture on new classes while utilizing pseudo labeling for the old classes. No additional distillation techniques are employed. The image-aware organ-specific heads, designed to be lightweight, are an integral part of our architecture. Each head comprises three convolution layers, with the first two layers having 8 kernels and the last layer having 1 kernel. For all compared models, we employ the AdamW optimizer and train them for 100 epochs using a cosine learning rate scheduler. During training, we utilize a batch size of 2 and a patch size of $96 \times 96 \times 96$. The initial learning rate is set to $1e^{-4}$, and the weight decay is set to $1e^{-5}$. We conduct our experiments using MONAI version 1.1.0. The training process is performed on NVIDIA TITAN RTX and Quadro RTX 8000 GPUs.

5.2. Results for extension to novel classes

The continual segmentation results for the tumor extension benchmark and body extension benchmark are presented in Tables 10 and 11, respectively. Notably, our model achieves the highest DSC score of 81.7% and 46.6% for all organs and tumors during step 2, as shown in Table 10. Furthermore, our model outperforms other methods in the body extension benchmark, indicating its superior ability to adapt to new data and classes while minimizing forgetting of previously learned classes.

To provide a qualitative assessment, we display the segmentation results of our proposed method alongside the best baseline method, ILT, for the body extension benchmark in Fig. 6. For instance, in case 2, ILT exhibits confusion between the stomach and pancreas. In contrast, our model accurately distinguishes these two organs due to the independent prediction enabled by the proposed LPG and CSH techniques, which prevent contradictions between different organs. The visualization effectively demonstrates that our method consistently achieves correct organ segmentation.

5.3. Ablation study

To assess the efficacy of the proposed model designs, we conducted an ablation study on the JHH dataset, as presented in Table 11. Specifically, we examined the performance impact of the organ-specific segmentation heads and the CLIP text embeddings. The first row in Table 11 corresponds to the performance of the baseline Swin UNETR model trained with pseudo labeling (LwF). The fourth row (Ours_1-hot) introduces the organ-specific segmentation heads but utilizes one-hot embeddings for each organ instead of the CLIP text embeddings. The third row represents the performance of the full method, incorporating both the organ-specific segmentation heads and the CLIP text embeddings. The results demonstrate a substantial improvement by adopting organ-specific heads as segmentation outputs. For instance, in step 2, we observe an improvement of 14.4% and a remarkable 17.9% enhancement in step 3 for gastrointestinal tracts. Furthermore, the utilization of CLIP text embeddings leads to further performance gains, with a margin of 18.4% improvement in step 3 for the cardiovascular system. These findings validate the effectiveness of the proposed organ-specific segmentation heads and CLIP text embeddings in the continual organ segmentation task.

Table 10
Results in tumor extension benchmark. We present the DSC score of each class in two continual learning steps. The Organ Mean represents the average DSC score for 13 organs, and Mean stands for the average score for 14 classes. The average performance is statistically significant at the P = 0.05 level, with highlighting in a light red box.

	Spleen		R kidney		L kidney		Gall Bladder		Esophagus		Liver		Stomach		Aorta	
	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2
LwF		0.941		0.745		0.845		0.796		0.690		0.935		0.852		0.847
ILT	0.955	**0.950**	0.819	0.687	0.927	**0.913**	0.844	0.750	0.718	**0.692**	0.969	**0.960**	0.896	0.857	0.801	0.774
PLOP		0.942		0.778		0.908		**0.823**		0.690		0.959		0.883		0.803
Ours	0.952	0.941	0.917	**0.860**	0.922	0.872	0.840	0.728	0.720	0.690	0.966	0.955	0.886	**0.862**	0.903	**0.909**
	IVC		Veins		Pancreas		R gland		L gland		Liver Tumor		Organ Mean		Mean	
	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2	Step 1	Step 2
LwF		0.784		0.628		0.812		0.671		0.468		0.456		0.770		0.748
ILT	0.877	0.849	0.735	**0.757**	0.826	0.812	0.743	0.692	0.654	0.527	N/A	0.335	0.828	0.786	N/A	0.754
PLOP		0.819		0.634		**0.822**		**0.708**		0.612		0.362		0.799		0.768
Ours	0.902	**0.872**	0.812	0.742	0.845	0.808	0.755	0.703	0.758	**0.683**	N/A	**0.466**	0.860	**0.817**	N/A	**0.792**

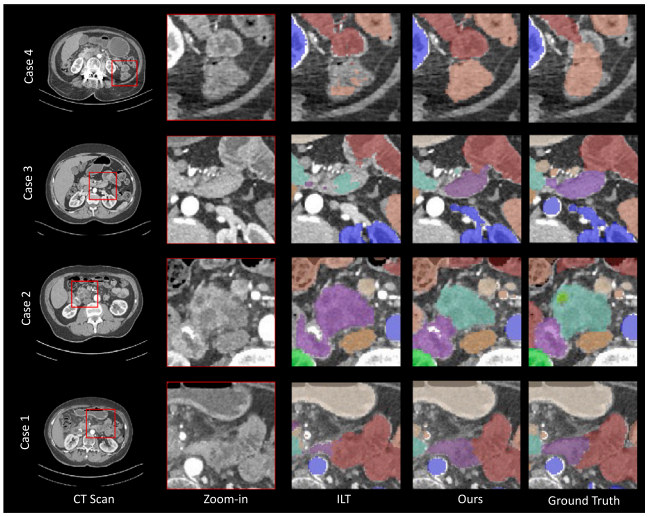


Fig. 6. *The visualization for body extension benchmark.* We compare Continual Universal Model with baseline model ILT across four case studies. Our model demonstrates effective adaptation to new organs while retaining knowledge of previously learned classes, thereby avoiding forgetting.

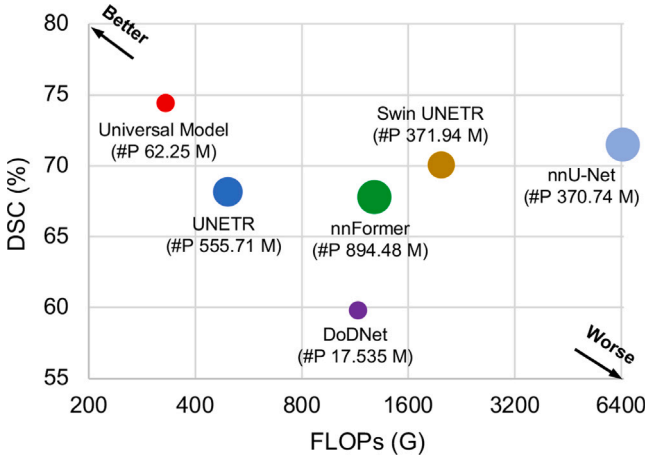


Fig. 7. *Inference speed with FLOPs vs. DSC.* The plot illustrates the relationship between the average DSC similarity coefficient (DSC) score across the six Medical Segmentation Decathlon (MSD) tasks and the floating-point operations per second (FLOPs). The FLOPs are calculated based on input with a spatial size of 96 × 96 × 96. The size of each circle in the plot corresponds to the number of parameters ('#P'). The used backbone of Universal Model is Swin UNETR.

5.4. Computational complexity analysis

Our work significantly reduces computational complexity in continual learning for segmentation. Ji et al. (2023) introduced a state-of-the-art architecture for medical continual semantic segmentation that uses a pre-trained, frozen encoder with incrementally added decoders at each learning step. This approach, however, leads to substantial computational complexity in subsequent steps. For example, the Swin UNETR's decoder alone requires 466.08 GFLOPs, adding this amount with each new learning step. In contrast, our model adds only a few image-aware, organ-specific heads for new classes, each consuming just 9.44×10^{-4} GFLOPs. Consequently, the computational complexity of

Table 11
Results in body extension benchmark and ablation study. We present the average DSC score for three body parts in each step separately. The Ours_1-hot replace the language embedding with one-hot embedding.

Method	JHH_organ (7)			JHH_gastro (3)		JHH_cardiac (4)
	Step 1	Step 2	Step 3	Step 2	Step 3	Step 3
LwF	**0.891**	0.777	0.767	0.530	0.486	0.360
ILT	**0.891**	0.775	0.776	0.653	0.480	0.484
PLOP	**0.891**	0.780	0.777	0.427	0.464	0.318
Ours (1-hot)	0.882	0.767	0.777	0.674	0.665	0.452
Ours (CLIP)	0.887	**0.783**	**0.787**	**0.695**	**0.692**	**0.636**

our model remains nearly constant throughout the continual learning process, whereas Ji et al. experiences linear growth in computational complexity with each step.

Moreover, we also assessed the inference speed of various models, as this is crucial for clinical applications (Chen et al., 2019a; Esteva et al., 2021). We utilized floating-point operations per second (FLOPS) as a metric to measure the inference speed. Fig. 7 presents a speed-performance plot, demonstrating that our Universal Model achieves computational efficiency surpassing that of dataset-specific models, with a speed improvement of over 6 times. Remarkably, this efficiency is achieved while maintaining the highest average Dice similarity coefficient (DSC) score of 74%.

6. Conclusion and discussion

In this work, we present a novel framework for multi-organ segmentation and tumor detection. This work integrates language embedding with segmentation models to enable a flexible and powerful segmentor, allowing the model to learn from an integrated dataset for high-performing organ segmentation and tumor detection, ranking first in both MSD and BTCV. More importantly, we demonstrate that language embedding can establish the anatomical relationship. Finally, the experiment results validate several clinically important merits of the CLIP-Driven Universal Model, such as compelling efficiency, generalizability, transferability, and extensibility.

While the universal model framework proposed in this study is primarily designed and evaluated using CT volumes, the general principles and techniques introduced herein could potentially be extended to other medical imaging modalities, such as magnetic resonance imaging (MRI) and ultrasound, with appropriate adaptations. MRI and ultrasound imaging exhibit unique characteristics, including bias fields, speckle noise, low contrast, and anisotropic resolution, which would necessitate specific preprocessing techniques and data augmentation strategies tailored to these modalities. Moreover, the backbone of our model is flexible and capable of employing any 2D or 3D network architecture, rendering it suitable for other medical imaging modalities. Furthermore, the language-driven parameter generator (LPG) and class-specific segment heads (CSH) leverage language embeddings from large language models like CLIP, which have been trained on diverse data, including medical images and text. Consequently, these language models could potentially capture semantic relationships between organs and abnormalities across various medical imaging modalities.

Besides, there are some limitations of current works. We also propose some potential future directions for each limitation. (1) Multi-modal Integration: The dataset primarily includes abdominal CT volumes. Future models could benefit from integrating diverse imaging modalities like CT, MRI, and ultrasound into a unified framework. This integration could enhance the model's robustness and provide more comprehensive anatomical information for various clinical applications. (2) Language Embedding: The language embeddings used are derived from pre-trained models like CLIP, which are not optimized for the medical domain. Developing domain-specific language models tailored for medical imaging tasks could improve semantic encoding and overall model performance. (3) Unlabeled Region Utilization: Semi-supervised learning techniques, such as consistency loss and pseudo-labeling, could be incorporated to leverage hidden information in unlabeled regions. Integrating these techniques could improve the model's ability to learn from both labeled and unlabeled data, thereby enhancing its generalization capabilities.

CRediT authorship contribution statement

Jie Liu: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Yixiao Zhang:** Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Kang Wang:** Resources, Validation, Visualization. **Mehmet Can Yavuz:** Validation, Visualization. **Xiaoxi Chen:** Formal analysis, Investigation, Validation, Visualization, Writing – original draft. **Yixuan Yuan:** Supervision, Resources, Investigation, Funding acquisition. **Haoliang Li:** Supervision, Resources, Investigation, Funding acquisition. **Yang Yang:** Data curation, Resources, Supervision, Validation. **Alan Yuille:** Writing – review & editing, Supervision, Resources, Investigation, Funding acquisition. **Yucheng Tang:** Writing – review & editing, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Zongwei Zhou:** Writing – review & editing, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Zongwei Zhou reports financial support was provided by Johns Hopkins University. Zongwei Zhou reports a relationship with Johns Hopkins University that includes: employment and funding grants. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research, the Patrick J. McGovern Foundation Award, and the National Natural Science Foundation of China (62001410). We thank Ali Hatamizadeh, Jieneng Chen, Junfei Xiao, Yongyi Lu, and Wenxuan Li for their constructive suggestions at several stages of the project.

References

Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., van Ginneken, B., et al., 2021. The medical segmentation decathlon. arXiv preprint arXiv:2106.05735.

Bai, X., Xia, Y., 2022. An end-to-end framework for universal lesion detection with missing annotations. In: 2022 16th IEEE International Conference on Signal Processing. ICSP, IEEE, pp. 411–415.

Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al., 2023. The liver tumor segmentation benchmark (lits). Med. Image Anal. 84, 102680.

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

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901.

Cai, J., Xia, Y., Yang, D., Xu, D., Yang, L., Roth, H., 2019. End-to-end adversarial shape learning for abdomen organ deep segmentation. In: Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings 10. Springer, pp. 124–132.

Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., et al., 2022. Monai: An open-source framework for deep learning in healthcare. arXiv–2211.

Chambon, P., Bluethgen, C., Langlotz, C.P., Chaudhari, A., 2022. Adapting pretrained vision-language foundational models to medical imaging domains. arXiv preprint arXiv:2210.04133.

Chen, Q., Chen, X., Song, H., Xiong, Z., Yuille, A., Wei, C., Zhou, Z., 2024b. Towards generalizable tumor synthesis. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. URL: https://github.com/MrGiovanni/DiffTumor.

Chen, P.H.C., Gadepalli, K., MacDonald, R., Liu, Y., Kadowaki, S., Nagpal, K., Kohlberger, T., Dean, J., Corrado, G.S., Hipp, J.D., et al., 2019a. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. Nat. Med. 25, 1453–1457.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y., 2021a. Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Chen, S., Ma, K., Zheng, Y., 2019b. Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625.

Chen, K., Qin, T., Lee, V.H.F., Yan, H., Li, H., 2024a. Learning robust shape regularization for generalizable medical image segmentation. IEEE Trans. Med. Imaging.

Chen, X., Sun, S., Bai, N., Han, K., Liu, Q., Yao, S., Tang, H., Zhang, C., Lu, Z., Huang, Q., et al., 2021b. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. Radiother. Oncol. 160, 175–184.

Chen, J., Xia, Y., Yao, J., Yan, K., Zhang, J., Lu, L., Wang, F., Zhou, B., Qiu, M., Yu, Q., et al., 2023a. Cancerunit: Towards a single unified model for effective detection, segmentation, and diagnosis of eight major cancers using a large collection of ct scans. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21327–21338.

Chen, Q., Xu, J., Koltun, V., 2017. Fast image processing with fully-convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2497–2506.

Chen, X., Zheng, H., Li, Y., Ma, Y., Ma, L., Li, H., Fan, Y., 2023b. Versatile medical image segmentation learned from multi-source datasets via model self-disambiguation. arXiv preprint arXiv:2311.10696.

Conneau, A., Lample, G., 2019. Cross-lingual language model pretraining. Adv. Neural Inf. Process. Syst. 32.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dmitriev, K., Kaufman, A.E., 2019. Learning multi-class segmentations from single-class datasets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9501–9511.

Douillard, A., Chen, Y., Dapogny, A., Cord, M., 2021. Plop: Learning without forgetting for continual semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4040–4050.

Eslami, S., Meinel, C., De Melo, G., 2023. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In: Findings of the Association for Computational Linguistics: EACL 2023. pp. 1151–1163.

Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., Socher, R., 2021. Deep learning-enabled medical computer vision. NPJ Digit. Med. 4, 1–9.

Fang, X., Yan, P., 2020. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. IEEE Trans. Med. Imaging 39, 3619–3629.

Gao, Y., Huang, R., Yang, Y., Zhang, J., Shao, K., Tao, C., Chen, Y., Metaxas, D.N., Li, H., Chen, M., 2021. Focusnetv2: Imbalanced large and small organ segmentation with adversarial shape constraint for head and neck ct images. Med. Image Anal. 67, 101831.

Gao, Y., Li, Z., Liu, D., Zhou, M., Zhang, S., Meta, D.N., 2023. Training like a medical resident: Universal medical image segmentation via context prior learning. arXiv preprint arXiv:2306.02416.

Germain, T., Favelier, S., Cercueil, J.P., Denys, A., Krausé, D., Guiu, B., 2014. Liver segmentation: practical tips. Diagn. Interv. Imaging 95, 1003–1016.

Guo, X., Liu, J., Yuan, Y., 2021b. Semantic-oriented labeled-to-unlabeled distribution translation for image segmentation. IEEE Trans. Med. Imaging 41, 434–445.

Guo, P., Wang, P., Zhou, J., Jiang, S., Patel, V.M., 2021a. Multi-institutional collaborations for improving deep learning-based magnetic resonance image reconstruction using federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2423–2432.

Haghighi, F., Taher, M.R.H., Zhou, Z., Gotway, M.B., Liang, J., 2021. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. IEEE Trans. Med. Imaging URL: https://github.com/fhaghighi/SemanticGenesis.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D., 2022a. 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. Springer, pp. 272–284.

Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D., 2022b. Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 574–584.

He, Y., Nath, V., Yang, D., Tang, Y., Myronenko, A., Xu, D., 2023. Swinunetr-v2: Stronger swin transformers with stagewise convolutions for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 416–426.

He, Y., Yang, D., Roth, H., Zhao, C., Xu, D., 2021. Dints: Differentiable neural network topology search for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5841–5850.

Heller, N., McSweeney, S., Peterson, M.T., Peterson, S., Rickman, J., Stai, B., Tejpaul, R., Oestreich, M., Blake, P., Rosenberg, J., et al., 2020. An international challenge to use artificial intelligence to define the state-of-the-art in kidney and kidney tumor segmentation in ct imaging.

Heller, N., Sathianathen, N., Kalapara, A., Walczak, E., Moore, K., Kaluzniak, H., Rosenberg, J., Blake, P., Rengel, Z., Oestreich, M., et al., 2019. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. arXiv preprint arXiv:1904.00445.

Hu, Q., Chen, Y., Xiao, J., Sun, S., Chen, J., Yuille, A.L., Zhou, Z., 2023. Label-free liver tumor segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7422–7432, URL: https://github.com/MrGiovanni/SyntheticTumors.

Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., Wang, L., 2022. Scaling up vision-language pre-training for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17980–17989.

Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J., 2023. A visual–language foundation model for pathology image analysis using medical twitter. Nat. Med. 29, 2307–2316.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. Nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 18, 203–211.

Jaus, A., Seibold, C., Hermann, K., Walter, A., Giske, K., Haubold, J., Kleesiek, J., Stiefelhagen, R., 2023. Towards unifying anatomy segmentation: automated generation of a full-body ct dataset via knowledge aggregation and anatomical guidelines. arXiv preprint arXiv:2307.13375.

Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al., 2022. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In: Neural Information Processing Systems. NeurIPS.

Ji, Z., Guo, D., Wang, P., Yan, K., Lu, L., Xu, M., Wang, Q., Ge, J., Gao, M., Ye, X., et al., 2023. Continual segment: Towards a single, unified and non-forgetting continual segmentation model of 143 whole-body organs in ct scans. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21140–21151.

Ji, W., Yu, S., Wu, J., Ma, K., Bian, C., Bi, Q., Li, J., Liu, H., Cheng, L., Zheng, Y., 2021. Learning calibrated medical image segmentation via multi-rater agreement modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12341–12351.

Jiang, Y., Huang, Z., Zhang, R., Zhang, X., Zhang, S., 2023. Zept: Zero-shot pan-tumor segmentation via query-disentangling and self-prompting. arXiv preprint arXiv:2312.04964.

Kim, S., Kim, I., Lim, S., Baek, W., Kim, C., Cho, H., Yoon, B., Kim, T., 2019. Scalable neural architecture search for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 220–228.

Lai, Y., Chen, X., Wang, A., Yuille, A., Zhou, Z., 2024. From pixel to cancer: Cellular automata in computed tomography. arXiv preprint arXiv:2403.06459. URL: https://github.com/MrGiovanni/Pixel2Cancer.

Landman, B., Xu, Z., Eugenio Igelsias, J., Styner, M., Robin Langerak, T., Klein, A., 2017. Multi-atlas labeling beyond the cranial vault-workshop and challenge.

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

Lewandowsky, S., Li, S.C., 1995. Catastrophic interference in neural networks: Causes, solutions, and data. In: Interference and Inhibition in Cognition. Elsevier, pp. 329–361.

Li, B., Chou, Y.C., Sun, S., Qiao, H., Yuille, A., Zhou, Z., 2023. Early detection and localization of pancreatic cancer by label-free tumor synthesis. In: MICCAI Workshop on Big Task Small Data, 1001-AI. URL: https://github.com/MrGiovanni/SyntheticTumors.

Li, Z., Hoiem, D., 2017. Learning without forgetting. IEEE Trans. Pattern Anal. Mach. Intell. 40, 2935–2947.

Li, W., Yuille, A., Zhou, Z., 2024. How well do supervised models transfer to 3d image segmentation? In: International Conference on Learning Representations. URL: https://github.com/MrGiovanni/SuPreM.

Liang, X., Li, N., Zhang, Z., Xiong, J., Zhou, S., Xie, Y., 2021. Incorporating the hybrid deformable model for improving the performance of abdominal ct segmentation via multi-scale feature fusion network. Med. Image Anal. 73, 102156.

Liu, P., Deng, Y., Wang, C., Hui, Y., Li, Q., Li, J., Luo, S., Sun, M., Quan, Q., Yang, S., et al., 2022a. Universal segmentation of 33 anatomies. arXiv preprint arXiv:2203.02098.

Liu, P., Gu, C., Wu, B., Liao, X., Qian, Y., Chen, G., 2023b. 3D multi-organ and tumor segmentation based on re-parameterize diverse experts. Mathematics 11, 4868.

Liu, J., Guo, X., Yuan, Y., 2021. Graph-based surgical instrument adaptive segmentation via domain-common knowledge. IEEE Trans. Med. Imaging 41, 715–726.

Liu, Z., Han, K., Xue, K., Song, Y., Liu, L., Tang, Y., Zhu, Y., 2022c. Improving ct-image universal lesion detection with comprehensive data and feature enhancements. Multimedia Syst. 1–12.

Liu, P., Wang, X., Fan, M., Pan, H., Yin, M., Zhu, X., Du, D., Zhao, X., Xiao, L., Ding, L., et al., 2022b. Learning incrementally to segment multiple organs in a ct image. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part IV. Springer, pp. 714–724.

Liu, X., Wen, B., Yang, S., 2023c. Ccq: cross-class query network for partially labeled organ segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 1755–1763.

Liu, H., Xu, Z., Gao, R., Li, H., Wang, J., Chabin, G., Oguz, I., Grbic, S., 2024. Cosst: Multi-organ segmentation with partially labeled datasets using comprehensive supervisions and self-training. IEEE Trans. Med. Imaging.

Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., Landman, B.A., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z., 2023a. Clip-driven universal model for organ segmentation and tumor detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21152–21164, URL: https://github.com/ljwztc/CLIP-Driven-Universal-Model.

Lüddecke, T., Ecker, A., 2022. Image segmentation using text and image prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7086–7096.

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

Ma, J., Zhang, Y., Gu, S., Zhu, C., Ge, C., Zhang, Y., An, X., Wang, C., Wang, Q., Liu, X., et al., 2021. Abdomenct-1k: Is abdominal organ segmentation a solved problem. IEEE Trans. Pattern Anal. Mach. Intell..

Mahmood, F., Borders, D., Chen, R.J., McKay, G.N., Salimian, K.J., Baras, A., Durr, N.J., 2019. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. IEEE Trans. Med. Imaging 39, 3257–3267.

Mattikalli, T., Mathai, T.S., Summers, R.M., 2022. Universal lesion detection in ct scans using neural network ensembles. In: Medical Imaging 2022: Computer-Aided Diagnosis. SPIE, pp. 864–868.

Michieli, U., Zanuttigh, P., 2019. Incremental learning techniques for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops.

Mongan, J., Moy, L., Kahn, C.E., 2020. Checklist for artificial intelligence in medical imaging (claim): A guide for authors and reviewers. Radiol.: Artif. Intell. 2, e200029. http://dx.doi.org/10.1148/ryai.2020200029, arXiv:https://doi.org/10.1148/ryai.2020200029. pMID: 33937821.

Myronenko, A., 2019. 3D mri brain tumor segmentation using autoencoder regularization. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4. Springer, pp. 311–320.

Naga, V., Mathai, T.S., Paul, A., Summers, R.M., 2022. Universal lesion detection and classification using limited data and weakly-supervised self-training. In: Workshop on Medical Image Learning with Limited and Noisy Data. Springer, pp. 55–64.

Norgeot, B., Quer, G., Beaulieu-Jones, B.K., Torkamani, A., Dias, R., Gianfrancesco, M., Arnaout, R., Kohane, I.S., Saria, S., Topol, E., et al., 2020. Minimum information about clinical artificial intelligence modeling: the mi-claim checklist. Nat. Med. 26, 1320–1324.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

Orbes-Arteaga, M., Varsavsky, T., Sudre, C.H., Eaton-Rosen, Z., Haddow, L.J., Sørensen, L., Nielsen, M., Pai, A., Ourselin, S., Modat, M., et al., 2019. Multi-domain adaptation in brain mri through paired consistency and adversarial learning. In: Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data. Springer, pp. 54–62.

Ozdemir, F., Fuernstahl, P., Goksel, O., 2018. Learn the new, keep the old: Extending pretrained models with new anatomy and images. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11. Springer, pp. 361–369.

Ozdemir, F., Goksel, O., 2019. Extending pretrained segmentation networks with additional anatomical structures. Int. J. Comput. Assist. Radiol. Surg. 14, 1187–1195.

Park, K., Woo, S., Oh, S.W., Kweon, I.S., Lee, J.Y., 2022. Per-clip video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1352–1361.

Qin, Z., Yi, H.H., Lao, Q., Li, K., 2022. Medical image understanding with pretrained vision language models: A comprehensive study. In: The Eleventh International Conference on Learning Representations.

Qu, C., Zhang, T., Qiao, H., Liu, J., Tang, Y., Yuille, A., Zhou, Z., 2023. Abdomenatlas-8k: Annotating 8,000 abdominal ct volumes for multi-organ segmentation in three weeks. In: Conference on Neural Information Processing Systems. URL: https://github.com/MrGiovanni/AbdomenAtlas.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR, pp. 8748–8763.

Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J., 2022. Denseclip: Language-guided dense prediction with context-aware prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18082–18091.

Rister, B., Yi, D., Shivakumar, K., Nobashi, T., Rubin, D.L., 2020. Ct-org, a new dataset for multiple organ segmentation in computed tomography. Sci. Data 7, 1–9.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Roth, H.R., Lu, L., Farag, A., Shin, H.C., Liu, J., Turkbey, E.B., Summers, R.M., 2015. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 556–564.

Schoppe, O., Pan, C., Coronel, J., Mai, H., Rong, Z., Todorov, M.I., Müskes, A., Navarro, F., Li, H., Ertürk, A., et al., 2020. Deep learning-enabled multi-organ segmentation in whole-body mouse scans. Nat. Commun. 11, 5626.

Shen, Y., Shamout, F.E., Oliver, J.R., Witowski, J., Kannan, K., Park, J., Wu, N., Huddleston, C., Wolfson, S., Millet, A., et al., 2021. Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. Nat. Commun. 12, 1–13.

Shi, G., Xiao, L., Chen, Y., Zhou, S.K., 2021. Marginal loss and exclusion loss for partially supervised multi-organ segmentation. Med. Image Anal. 70, 101979.

Siddiquee, M.M.R., Myronenko, A., 2021. Redundancy reduction in semantic segmentation of 3d brain tumor mris. arXiv preprint arXiv:2111.00742.

Silva-Rodríguez, J., Dolz, J., Ayed, I.B., 2023. Towards foundation models and few-shot parameter-efficient fine-tuning for volumetric organ segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 213–224.

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

Soler, L., Hostettler, A., Agnus, V., Charnoz, A., Fasquel, J., Moreau, J., Osswald, A., Bouhadjar, M., Marescaux, J., 2010. 3d Image Reconstruction for Comparison of Algorithm Database: A Patient Specific Anatomical and Medical Image Database. Tech. Rep, IRCAD, Strasbourg, France.

Tang, Y., Liu, J., Zhou, Z., Yu, X., Huo, Y., 2024. Efficient 3d representation learning for medical image analysis. World Sci. Annu. Rev. Artif. Intell..

Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A., 2022. Self-supervised pre-training of swin transformers for 3d medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20730–20740.

Ulrich, C., Isensee, F., Wald, T., Zenk, M., Baumgartner, M., Maier-Hein, K.H., 2023. Multitalent: A multi-dataset approach to medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 648–658.

Valindria, V.V., Pawlowski, N., Rajchl, M., Lavdas, I., Aboagye, E.O., Rockall, A.G., Rueckert, D., Glocker, B., 2018. Multi-modal learning from unpaired images: Application to multi-organ segmentation in ct and mri. In: 2018 IEEE Winter Conference on Applications of Computer Vision. WACV, IEEE, pp. 547–556.

Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., Li, J., 2021. Transbts: Multimodal brain tumor segmentation using transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 109–119.

Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T., 2022a. Cris: Clip-driven referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11686–11695.

Wang, Z., Wu, Z., Agarwal, D., Sun, J., 2022b. Medclip: Contrastive learning from unpaired medical images and text. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 3876–3887.

Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al., 2023. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiol.: Artif. Intell. 5.

Wasserthal, J., Meyer, M., Breit, H.C., Cyriac, J., Yang, S., Segeroth, M., 2022. Totalsegmentator: robust segmentation of 104 anatomical structures in ct images. arXiv preprint arXiv:2208.05868.

Wu, H., Pang, S., Sowmya, A., 2022. Tgnet: A task-guided network architecture for multi-organ and tumour segmentation from partially labelled datasets. In: 2022 IEEE 19th International Symposium on Biomedical Imaging. ISBI, IEEE, pp. 1–5.

Xia, Y., Yu, Q., Chu, L., Kawamoto, S., Park, S., Liu, F., Chen, J., Zhu, Z., Li, B., Zhou, Z., et al., 2022. The felix project: Deep networks to detect pancreatic neoplasms. medRxiv.

Xie, J., Hou, X., Ye, K., Shen, L., 2022a. Clims: Cross language image matching for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4483–4492.

Xie, Y., Zhang, J., Shen, C., Xia, Y., 2021. Cotr: Efficiently bridging cnn and transformer for 3d 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. Springer, pp. 171–180.

Xie, Y., Zhang, J., Xia, Y., Shen, C., 2023. Learning from partially labeled data for multi-organ and tumor segmentation. IEEE Trans. Pattern Anal. Mach. Intell..

Xie, Y., Zhang, J., Xia, Y., Wu, Q., 2022b. Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In: European Conference on Computer Vision. Springer, pp. 558–575.

Yan, K., Cai, J., Harrison, A.P., Jin, D., Xiao, J., Lu, L., 2020a. Universal lesion detection by learning from multiple heterogeneously labeled datasets. arXiv preprint arXiv:2005.13753.

Yan, W., Huang, L., Xia, L., Gu, S., Yan, F., Wang, Y., Tao, Q., 2020b. Mri manufacturer shift and adaptation: increasing the generalizability of deep learning segmentation for mr images acquired with different scanners. Radiol.: Artif. Intell. 2, e190195.

Yan, B., Pei, M., 2022. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 2982–2990.

Yan, K., Tang, Y., Peng, Y., Sandfort, V., Bagheri, M., Lu, Z., Summers, R.M., 2019. Mulan: multitask universal lesion analysis network for joint lesion detection, tagging, and segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 194–202.

Yasunaga, M., Leskovec, J., Liang, P., 2022. Linkbert: Pretraining language models with document links. arXiv preprint arXiv:2203.15827.

Ye, Y., Xie, Y., Zhang, J., Chen, Z., Wu, Q., Xia, Y., 2023a. Continual self-supervised learning: Towards universal multi-modal medical data representation learning. arXiv preprint arXiv:2311.17597.

Ye, Y., Xie, Y., Zhang, J., Chen, Z., Xia, Y., 2023b. Uniseg: A prompt-driven universal segmentation model as well as a strong representation learner. arXiv preprint arXiv:2304.03493.

Yu, Q., Yang, D., Roth, H., Bai, Y., Zhang, Y., Yuille, A.L., Xu, D., 2020. C2fnas: Coarse-to-fine neural architecture search for 3d medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4126–4135.

Yu, X., Yang, Q., Zhou, Y., Cai, L.Y., Gao, R., Lee, H.H., Li, T., Bao, S., Xu, Z., Lasko, T.A., et al., 2022. Unest: Local spatial representation learning with hierarchical transformer for efficient medical segmentation. arXiv preprint arXiv:2209.14378.

Zeng, Q., Xie, Y., Lu, Z., Lu, M., Wu, Y., Xia, Y., 2023. Segment together: A versatile paradigm for semi-supervised medical image segmentation. arXiv preprint arXiv:2311.11686.

Zhang, T., Chen, X., Qu, C., Yuille, A., Zhou, Z., 2024. Leveraging ai predicted and expert revised annotations in interactive segmentation: Continual tuning or full training? In: IEEE International Symposium on Biomedical Imaging. IEEE, URL: https://github.com/MrGiovanni/ContinualLearning.

Zhang, H., Chen, X., Wang, R., Hu, R., Liu, D., Li, G., 2023a. Spatially covariant image registration with text prompts. arXiv preprint arXiv:2311.15607.

Zhang, Y., Li, X., Chen, H., Yuille, A.L., Liu, Y., Zhou, Z., 2023b. Continual learning for abdominal multi-organ and tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 35–45, URL: https://github.com/MrGiovanni/ContinualLearning.

Zhang, J., Xie, Y., Xia, Y., Shen, C., 2021. 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. pp. 1195–1204.

Zhang, W., Zhang, J., Wang, X., Yang, S., Huang, J., Yang, W., Wang, W., Han, X., 2022. Merging nucleus datasets by correlation-based cross-training. Med. Image Anal. 102705.

Zhou, Z., 2021. Towards Annotation-Efficient Deep Learning for Computer-Aided Diagnosis (Ph.D. thesis). Arizona State University.

Zhou, Z., Gotway, M.B., Liang, J., 2022. Interpreting medical images. In: Intelligent Systems in Medicine and Health. Springer, pp. 343–371.

Zhou, H.Y., Guo, J., Zhang, Y., Yu, L., Wang, L., Yu, Y., 2021a. Nnformer: Interleaved transformer for volumetric segmentation. arXiv preprint arXiv:2109.03201.

Zhou, Y., Li, Z., Bai, S., Wang, C., Chen, X., Han, M., Fishman, E., Yuille, A.L., 2019a. Prior-aware neural network for partially-supervised multi-organ segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10672–10681.

Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2019b. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans. Med. Imaging 39, 1856–1867, URL: https://github.com/MrGiovanni/UNetPlusPlus.

Zhou, Z., Sodha, V., Pang, J., Gotway, M.B., Liang, J., 2021b. Models genesis. Med. Image Anal. 67, 101840, URL: https://github.com/MrGiovanni/ModelsGenesis.

Zhou, Z., Sodha, V., Siddiquee, M.M.R., Feng, R., Tajbakhsh, N., Gotway, M.B., Liang, J., 2019c. Models genesis: Generic autodidactic models for 3d medical image analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 384–393, URL: https://github.com/MrGiovanni/ModelsGenesis.

Zlocha, M., Glocker, B., Passerat-Palmbach, J., 2019. Universal lesion detector: Deep learning for analysing medical scans.