CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection

Jie Liu[1], Yixiao Zhang[2], Jie-Neng Chen[2], Junfei Xiao[2], Yongyi Lu[2], Bennett A. Landman[3], Yixuan Yuan[4,5], Alan Yuille[2], Yucheng Tang[3,6,*], and Zongwei Zhou[2,*]

[1]City University of Hong Kong [2]Johns Hopkins University [3]Vanderbilt University
[4]Chinese University of Hong Kong [5]CUHK Shenzhen Research Institute [6]NVIDIA

Project: https://github.com/ljwztc/CLIP-Driven-Universal-Model

Abstract

An increasing number of public datasets have shown a marked impact on automated organ segmentation and tumor detection. However, due to the small size and partially labeled problem of each dataset, as well as a limited investigation of diverse types of tumors, the resulting models are often limited to segmenting specific organs/tumors and ignore the semantics of anatomical structures, nor can they be extended to novel domains. To address these issues, we propose the CLIP-Driven Universal Model, which incorporates text embedding learned from Contrastive Language-Image Pre-training (CLIP) to segmentation models. This CLIP-based label encoding captures anatomical relationships, enabling the model to learn a structured feature embedding and segment 25 organs and 6 types of tumors. The proposed model is developed from an assembly of 14 datasets, using a total of 3,410 CT scans for training and then evaluated on 6,162 external CT scans from 3 additional datasets. We rank first on the Medical Segmentation Decathlon (MSD) public leaderboard and achieve state-of-the-art results on Beyond The Cranial Vault (BTCV). Additionally, the Universal Model is computationally more efficient (6× faster) compared with dataset-specific models, generalized better to CT scans from varying sites, and shows stronger transfer learning performance on novel tasks.

1. Introduction

Enormous advances in medical imaging benefit from the ever-growing number of annotated datasets [43, 1, 42, 31, 76]. Although a total of around 5,000 annotated abdominal CT scans are publicly available, it is still commonly perceived that medical imaging datasets are too small to develop robust AI models [94, 72, 54, 66, 95, 13]. One reason for this impression is the high cost of detailed per-voxel segmentation annotations, which can take nearly one hour per organ for an expert annotator. Since each institute has time, monetary, and clinical constraints, the number of CT scans in each dataset is limited, and the types of annotated organs vary significantly from institute to institute. Moreover, only a small proportion (hundreds) of public CT scans contain tumor annotation performed by experts [3, 24, 1].

The partially labeled problem [32, 88, 37] can impose significant limitations on the performance of models trained on existing public datasets, ultimately hindering their effectiveness for multi-organ segmentation and tumor detection. However, despite this challenge, the potential of AI models in these areas remains promising and largely unexplored. This has motivated us to exploit the public datasets with partial labels, and demonstrate the clinical impact of AI framework, including model expansibility (*i.e.*, adaptable to various network backbone), generalizability (*i.e.*, robust to CT scans from various hospitals) [43] and transferability (*i.e.*, generic image representation that is transferable to multiple

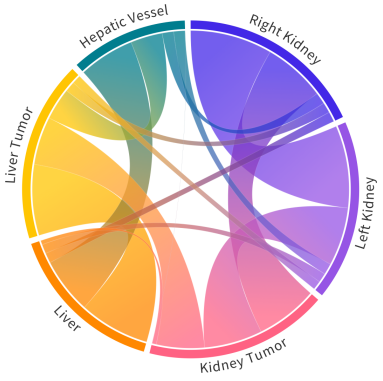


Figure 1. **Cosine similarity between CLIP embeddings.** The CLIP embedding reveals the intrinsic semantics of the anatomical structures by mapping similar concepts close to each other in the embedding space. For example, "Liver" has a large similarity with "Liver Tumor" and "Hepatic Vessel" (the hepatic vessel returns low-oxygen blood from the liver to the heart, which has a high anatomical relationship with the liver).

*Corresponding authors: Yucheng Tang (yuchengt@nvidia.com) and Zongwei Zhou (zzhou82@jh.edu)

downstream tasks) [97]. Specifically, we have assembled 14 publicly available datasets, including 3,410 CT scans with 25 partially annotated organs and 6 tumors.

Formidable challenges exist in assembling partially annotated datasets. **First,** label inconsistency, in five aspects. *(i)* Index inconsistency. The same organ can be labeled as different indexes. For example, the stomach is labeled '7' in BTCV, but '5' in WORD. *(ii)* Name inconsistency. Naming can be confusing if multiple labels refer to the same anatomical structure. For example, "postcava" in AMOS22 and "inferior vena cava" in BTCV. *(iii)* Background inconsistency. For example, when combining Pancreas-CT and MSD-Spleen, the pancreas is marked as the background in MSD-Spleen, but it should have been marked as the foreground. *(iv)* Organ overlapping. There is overlap between various organs. For example, "Hepatic Vessel" is part of the "Liver" and "Kidney Tumor" is a sub-volume of the "Kidney". *(v)* Data overlapping. Some CT scans are overlapped among public datasets, but with different annotations. For example, KiTS is part of AbdomenCT-1K, and kidney tumor is annotated in KiTS rather than AbdomenCT-1K. **Second,** label orthogonality. Most segmentation methods, trained with one-hot labels [88], ignore the semantic relationship between classes. Given one-hot labels of liver [1,0,0], liver tumor [0,1,0], and pancreas [0,0,1], there is no semantic difference between liver↔liver tumor and liver↔pancreas. A possible solution is few-hot labels [62], with which, the liver, liver tumor, and pancreas can be encoded as [1,0,0], [1,1,0], and [0,0,1]. Although few-hot labels could indicate that liver tumors are part of the liver, the relationship between organs remains orthogonal.

To address above mentioned challenged, CLIP-driven *Universal Model* incorporates text embedding and adopts masked back-propagation mechanism with binary segmentation mask. Specifically, we maintain a revised label taxonomy derived from a collection of public datasets and generate a binary segmentation mask for each class during image pre-processing. For architecture design, we draw inspiration from Guo *et al*. [19] and replaced one- or few-hot labels with the text embedding generated by the pre-trained text encoder from CLIP[1]. Figure 1 illustrates how CLIP embedding presents the relationship between organs and tumors. This CLIP-based label encoding enhances the anatomical structure of universal model feature embedding, which is visualized in Figure 6. At last, we only compute loss for the classes with available labels.

In summary, this work proposes a CLIP-Driven *Universal Model* that allows superior segmentation of 25 organs and detection of 6 tumors with state-of-the-art performance. The Universal Model can be generalized to CT scans from different institutes. Experimental results have demonstrated **six advantages** of the CLIP-Driven Universal Model:

1. High abdominal organ segmentation performance. We rank first in the MSD and BTCV challenges, leading to substantial performance improvement. Moreover, six organs can be annotated by Universal Model with a similar intra-observer variability to humans.

2. Predicting fewer false positives than existing models while maintaining high sensitivity for tumor detection.

3. Computationally more efficient than dataset-specific models, accelerating the testing speed by factor of six.

4. The Universal Model framework can be expanded to various backbones such as CNNs and Transformers.

5. The performance of organ segmentation and tumor detection can be generalized to CT scans from a variety of hospitals without additional tuning and adaptation.

6. An effective Foundation Model for numerous downstream tasks, showing a strong transferability on tasks across multiple diseases, organs, and datasets.

2. Related Work

Partial label problem. Publicly available datasets for abdominal imaging focus on different organs and tumors [35, 43, 42, 31], e.g., AbdomenCT-1K dataset for 4 organ segmentation [43], WORD dataset for 16 organ segmentation [42] and TotalSegmentor dataset for 104 anatomical structure segmentation [76]. The partial label problem occurs when training AI models on a combination of these datasets due to their inconsistent label taxonomy. To exploit the partial labels, several approaches have been investigated [93, 18, 88, 89], aiming for a single model that can perform organ segmentation [39, 12] and tumor detection [2, 100, 80, 40, 48, 78, 45]. These studies have the following limitations. (1) Due to the small scale of the dataset assembly[2], the potential of assembling datasets was not convincing. Their performance was similar to dataset-specific models and was not evaluated on the official benchmark. (2) Due to the one-hot labels, the semantic relationship between organs and tumors was discarded. Table 1 reveals that the introduction of CLIP embedding is a salient factor to our proposed framework.

Organ segmentation and tumor detection. Deep learning-based methods have been widely applied to organ segmentation and tumor detection. U-Net [59] and its variants [96, 38, 51, 29] are one of the main streams and achieve

[1]CLIP (Contrastive Language–Image Pre-training) was pre-trained on 400 million image-text pairs (some are medical images and tex[t] ploiting the semantic relationship between images and language

[2]Zhou *et al*. [93] assembled 150 CT scans from 4 datasets; Fang *et* [8] assembled 548 CT scans from 4 datasets; Zhang *et al*. [88] assem- 1,155 CT scans from 7 datasets.

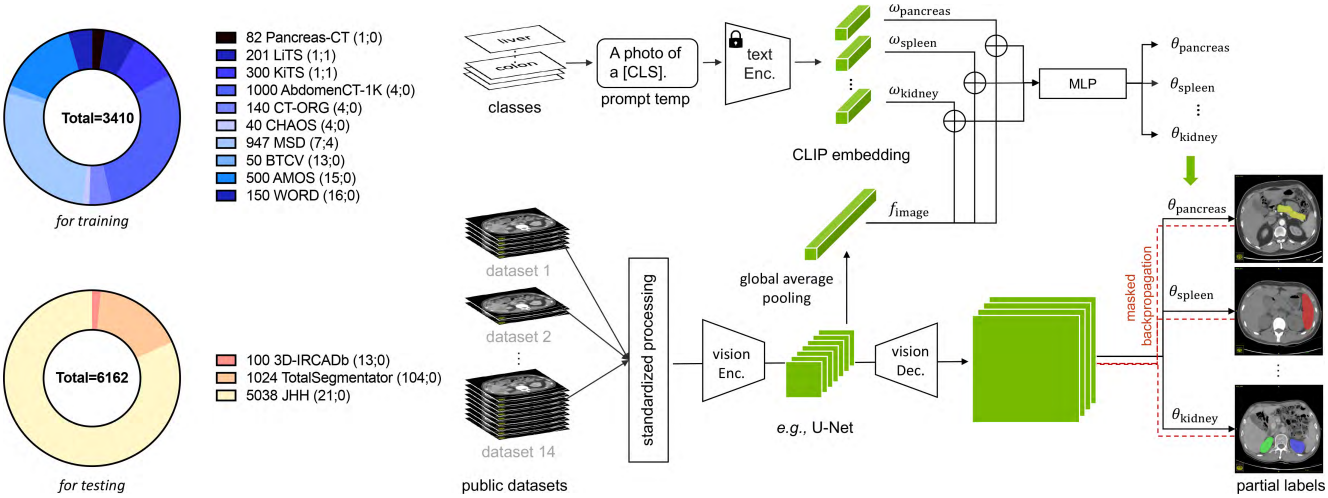


Figure 2. **Overview.** We have developed a Universal Model from an assembly of 14 public datasets of 3,410 CT scans. In total, 25 organs and 6 types of tumors are partially labeled (detailed in Appendix Table 7). To deal with partial labels, Universal Model consists of a text branch and a vision branch (§3.2). The official test set of MSD and BTCV are used to benchmark the performance of organ segmentation (§4.1) and tumor detection (§4.2). 3D-IRCADb, TotalSegmentator and a large-scale private dataset, consisting of 5,038 CT scans with 21 annotated organs, are used for independent, external validation of model generalizability and transferability (§5).

some promising results. Recently, transformer based models [8, 91, 22, 68, 7] are emerged, which can capture the global relationship between whole volume. These works are often specialized for single organ [59, 96, 29, 38] or single task, i.e., organ segmentation [91, 22, 68, 7] or tumor detection [8, 79, 81]. Different from these work, Universal Model tackles both tasks within a single framework, using the introduced CLIP embedding to capture the semantic relationship between organs and tumors. Moreover, we demonstrate our work on publically available datasets, which is beneficial to reproducibility.

CLIP in medical imaging. With the widespread success of large models in the field of language processing and understanding [15, 4, 64, 41], large-scale pre-trained vision-language models (VLM), *e.g.*, Conneau *et al.* [14], have recently been applied to multiple vision tasks [57, 74, 6, 53], but rarely to the medical domain [16, 75]. Qin *et al.* [55] suggested that VLM could be used for detection task in the medical domain with carefully designed medical prompts. Grounded in this findings, we are among the first to introduce CLIP embedding to voxel-level semantic understanding medical tasks, i.e., segmentation, in which we underline the importance of the semantic relationship between anatomical structures.

Medical universal models. The field of medical image analysis has undergone a significant shift from training individual models for specific datasets towards developing a single (universal/foundation) model that can effectively handle diverse datasets, organs, tumors, tasks, and modalities. After we first presented CLIP-Driven Universal Model in arXiv and released the code, the field has witnessed numer

otal contributions [85, 47, 70, 28, 84, 5], with many more endeavors underway to our knowledge [56, 90]. Thereby, we are dedicated to reviewing the exceptional studies in the field by actively maintaining a GitHub page.

3. Methodology

3.1. Background

Problem definition. Let M and N be the total number of datasets to combine and data points in the combination of the datasets, respectively. Given a dataset $\mathcal{D} = \{(\boldsymbol{X}_1, \boldsymbol{Y}_1), (\boldsymbol{X}_2, \boldsymbol{Y}_2), ..., (\boldsymbol{X}_N, \boldsymbol{Y}_N)\}$, there are a total of K unique classes. For $\forall n \in [1, N]$, if the presence of $\forall k \in [1, K]$ classes in $\boldsymbol{X}_i$ is annotated in $\boldsymbol{Y}_i$, $\mathcal{D}$ is a *fully labeled* dataset; otherwise, $\mathcal{D}$ is a *partially labeled* dataset.

Previous solutions. Two groups of solutions were proposed to address the partial label problem. Given a data point $\boldsymbol{X}_n, n \in [1, N]$, the objective is to train a model $\mathcal{F}(\cdot)$ using the assembly dataset $\mathcal{D}_A = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_M\}$, and the model can predict all K classes, if presented in $\boldsymbol{X}_n$.

- *Solution #1* [18, 62, 79, 62, 93, 11, 27, 68] aims to solve $\mathcal{F}_\theta(\boldsymbol{X}_n) = \boldsymbol{P}_n^k, n \in [1, N], k \in [1, K]$, where the prediction $\boldsymbol{P}_n$ is one-hot encoding with length k.

- *Solution #2* [88, 32, 99] aims to solve $\mathcal{F}_\theta(\boldsymbol{X}_n, \boldsymbol{w}_k) = \boldsymbol{P}_n, n \in [1, N], k \in [1, K]$, where $\boldsymbol{w}_k$ is an one-hot vector to indicate which class to be predicted.

According to Zhang *et al.* [88], both solutions have similar segmentation performance, but #2 is computationally

Table 1. **Label Encoding Ablation.** All three prompts can elicit knowledge from CLIP, achieving significant improvement over the conventional one-hot labels (DoDNet [88]) and BioBERT [83]. The average DSC score over validation part of Assembling Datasets is reported; per-class DSC found in Appendix Table 14.

Embedding	prompt	DSC
One-hot [88]	-	70.42
BioBERT [83]	A computerized tomography of a [CLS].	71.55
CLIP V1	A photo of a [CLS].	73.49
CLIP V2	There is [CLS] in this computerized tomography.	75.66
CLIP V3	A computerized tomography of a [CLS].	**76.11**

more efficient. However, both solutions rely on one-hot labels, sharing two limitations. First, they ignore the semantic and anatomical relationship between organs and tumors. Second, they are inappropriate for segmenting various subtypes of tumors. To address these limitations, we modify $\boldsymbol{w}_k$ in Solution #2 to CLIP embedding and introduce in-depth in the following sections.

3.2. CLIP-Driven Universal Model

The overall framework of CLIP-Driven Universal Model (see Figure 2) has a text branch and a vision branch. The text branch first generates the CLIP embedding for each organ and tumor using an appropriate medical prompting (Table 1), and then the vision branch takes both CT scans and CLIP embedding to predict the segmentation mask[3].

Text branch. Let $\boldsymbol{w}_k$ be the CLIP embedding of the k-th class, produced by the pre-trained text encoder in CLIP and a medical prompt (*e.g.*, "a computerized tomography of a [CLS]", where [CLS] is a concrete class name). We first concatenate the CLIP embedding ($\boldsymbol{w}_k$) and the global image feature ($\boldsymbol{f}$) and then input it to a multi-layer perceptron (MLP), namely *text-based controller* [69], to generate parameters ($\boldsymbol{\theta}_k$), *i.e.*, $\boldsymbol{\theta}_k = \text{MLP}(\boldsymbol{w}_k \oplus \boldsymbol{f})$, where $\oplus$ is the concatenation. Although CLIP embedding significantly outperforms one-hot labels [88], we mark that the choice of medical prompt template is critical. Table 1 presents the effectiveness of three prompt templates. Moreover, the introduction of CLIP embedding addresses the label orthogonality problem by exploiting semantic relationships among organs and tumors (illustrated in Figure 1).

Vision branch. We pre-process CT scans using isotropic spacing and uniformed intensity scale to reduce the domain gap among various datasets[4]. The standardized and normalized CT scans are then processed by the vision encoder. Let $\boldsymbol{F}$ be the image features extracted by the vision encoder. To process $\boldsymbol{F}$, we use three sequential convolutional layers with $1 \times 1 \times 1$ kernels, namely *text-driven segmentor*. The first two layers have 8 channels, and the last one has 1 channel, corresponding to the class of $[\text{CLS}]_k$. The prediction for the class $[\text{CLS}]_k$ is computed as $\boldsymbol{P}_k = \text{Sigmoid}\left(((\boldsymbol{F} * \boldsymbol{\theta}_{k_1}) * \boldsymbol{\theta}_{k_2}) * \boldsymbol{\theta}_{k_3}\right)$, where $\boldsymbol{\theta}_k = \{\boldsymbol{\theta}_{k_1}, \boldsymbol{\theta}_{k_2}, \boldsymbol{\theta}_{k_3}\}$ are computed in the text branch, and $*$ represents the convolution. For each class $[\text{CLS}]_k$, we generate the prediction $\boldsymbol{P}_k \in \mathbb{R}^{1\times D\times W\times H}$ representing the foreground of each class in *one vs. all* manner (*i.e.*, Sigmoid instead of Softmax).

Masked back-propagation. To address the label inconsistency problem, we proposed the masked back-propagation technique. The BCE loss function is utilized for supervision. We masked the loss terms of these classes that are not contained in $\boldsymbol{Y}$ and only back-propagate the accurate supervision to update the whole framework. The masked back-propagation addresses the label inconsistency in the partial label problem. Specifically, partially labeled datasets annotate some other organs as background, leading to the disability of existing training schemes (Solution #1).

4. Experiments & Results

Datasets and evaluation. 14 public datasets of 3,410 CT scans in total are assembled for training. Other two public and one private datasets are used for testing. Dataset details and pre-processing are in Appendix §B. Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD) are evaluated for organ/tumor segmentation; Sensitivity and Specificity are for tumor detection.

Implementation details. The Universal Model is trained using the AdamW optimizer with a warm-up cosine scheduler of 50 epochs. The segmentation experiments use batch-size of 6 per GPU with a patch size of $96 \times 96 \times 96$. Default initial learning rate of $4e^{-4}$, momentum of 0.9 and decay of $1e^{-5}$ on multi-GPU (4) with DDP. The framework is implemented in MONAI 0.9.0[5]. The five-fold cross validation strategy is performed. We select the best model in each fold by evaluating the validation best metrics. Models are trained on eight NVIDIA RTX A5000 cards.

4.1. Organ Segmentation on MSD and BTCV

We offer the top #1 solution in both Medical Segmentation Decathlon (MSD)[6] and Beyond The Cranial Vault (BTCV), surpassing the runners-up by a considerable margin. It's noted that universal model provides six CT tasks solution and the results of other four MRI tasks are predicted by nnUnet [29]. Table 2 and Figure 3 present de-

[3]Our framework design is conceptually similar to *Segment Anything Model (SAM)* [34], which is a concurrent study of ours in computer vision. By leveraging CLIP embedding as a prompt within our Universal Model, we are able to generate highly accurate masks for organs and tumors of interest, as opposed to producing masks for arbitrary objects.

[4]A standardized and normalized CT pre-processing is important when combining multiple datasets. Substantial differences in CT scans can occur in image quality and technical display, originating from different acquisition parameters, reconstruction kernels, contrast enhancements variation, and so on [52, 82, 20].

[5]https://monai.io/

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

Table 2. **Leaderboard performance on MSD.** The results are evaluated in the server on the MSD competition test dataset. All Dice and NSD metrics are obtained from the MSD public leaderboard. The results of MRI-related tasks were generated by Swin UNETR [68].

Method	Task03 Liver						Task07 Pancreas					
	Dice1	Dice2	Avg.	NSD1	NSD2	Avg.	Dice1	Dice2	Avg.	NSD1	NSD2	Avg.
Kim *et al*. [33]	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 [21]	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[86]	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. [97]	95.72	77.50	86.61	98.48	91.92	95.20	81.36	50.36	65.86	96.16	70.02	83.09
nnUNet [29]	**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 [23]	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 [68]	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	
	Dice1	Dice2	Avg.	NSD1	NSD2	Avg.	Dice1	NSD1	Dice1	NSD1	Dice1	NSD1
Kim *et al*. [33]	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 [21]	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[86]	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. [97]	65.80	71.44	68.62	84.01	80.15	82.08	74.54	76.22	97.35	99.87	51.47	60.53
nnUNet [29]	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 [23]	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 [68]	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**

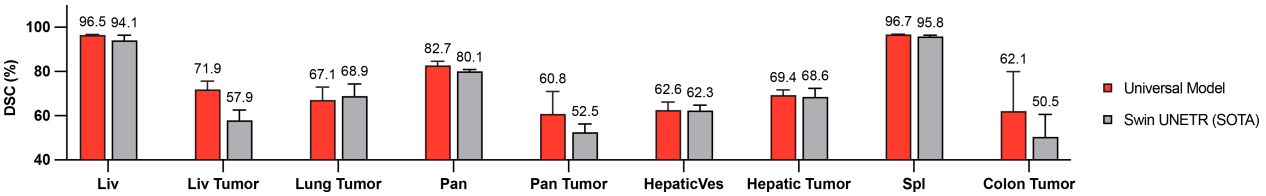


Figure 3. **Benchmark on MSD validation dataset.** We compare Universal Model with Swin UNETR [68] (previously ranked first on the MSD leaderboard) on 5-fold cross-validation of the MSD dataset. Universal Model achieves overall better segmentation performance and offers *substantial* improvement in the tasks of segmenting liver tumors (+14%), pancreatic tumors (+8%), and colon tumors (+11%).

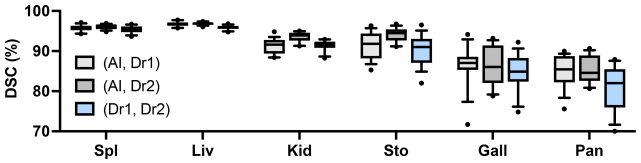


Figure 4. **Intra-observer variability.** We obtain similar performance between pseudo labels generated by the Universal Model (AI) and annotations performed by two human experts (Dr1,2) on 6 organs. Spleen (Spl), liver (Liv), kidneys (Kid), stomach (Sto), gallbladder (Gall), and pancreas (Pan) can be annotated by AI with a similar intra-observer variability to humans. Examples of pseudo labels and human annotations are provided in Appendix Figure 9.

tailed comparison on the official test set and 5-hold cross validation on MSD, respectively. Table 3 compares Universal Model with other methods in the validation set of BTCV, offering at least 3.5% improvements over the second best.

Manual annotations have inter-rater and intra-rater variance [30], particularly in segmentation tasks, because some of the organs' boundaries are blurry and ambiguous. We assess the quality of pseudo labels predicted by Universal Model and manual annotation performed by human experts. 17 CT scans in BTCV have been annotated by two i

dent groups of radiologists from different institutes (not test server labels). As a result, each CT scan is associated with AI prediction, and two human annotations (Dr1 and Dr2). Figure 4 presents their mutual DSC scores, *i.e.*, AI↔Dr1, AI↔Dr2, and Dr1↔Dr2. We find the DSC between AI and humans is slightly larger than the DSC between humans in segmenting 6 types of organs (*i.e.*, spleen, liver, kidney, stomach, and pancreas). With this high-quality AI prediction, we assemble a large dataset of 3,410 CT scans from a diverse set of hospitals (Figure 2 and generate pseudo labels for 25 organs and 6 tumors[7]. Pseudo-label refinement has been performed for a few CT scans where AI's prediction is uncertain. This fully annotated dataset will be released (examples in Appendix Figure 14). Now that these 6 organs can be segmented by AI with a similar variance to human experts, we encourage the research community to concentrate on creating annotations for harder organs and tumors.

4.2. Tumor Detection on Five Datasets

Figure 3 demonstrates that Universal Model surpasses Swin UNETR by a large margin in segmenting liver, pan-

[7]The quality of 19 other organs and 6 tumors has not been compared human annotations because there is no publicly available CT scans have been annotated by multiple independent groups on these objects.

Table 3. **5-fold cross-validation results on BTCV.** For a fair comparison, we did not use model ensemble during the evaluation. All experiments are under the same data splits, computing resources, and testing conditions. Universal Model achieves the overall best performance, yielding at least +3.9% DSC improvement over the state-of-the-art method.

Methods	Spl	RKid	LKid	Gall	Eso	Liv	Sto	Aor	IVC	Veins	Pan	AG	Avg.
RandPatch [67]	95.82	88.52	90.14	68.31	75.01	96.48	82.93	88.96	82.49	73.54	75.48	66.09	80.76
TransBTS [29]	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 [91]	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 [22]	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 [29]	**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 [68]	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
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**

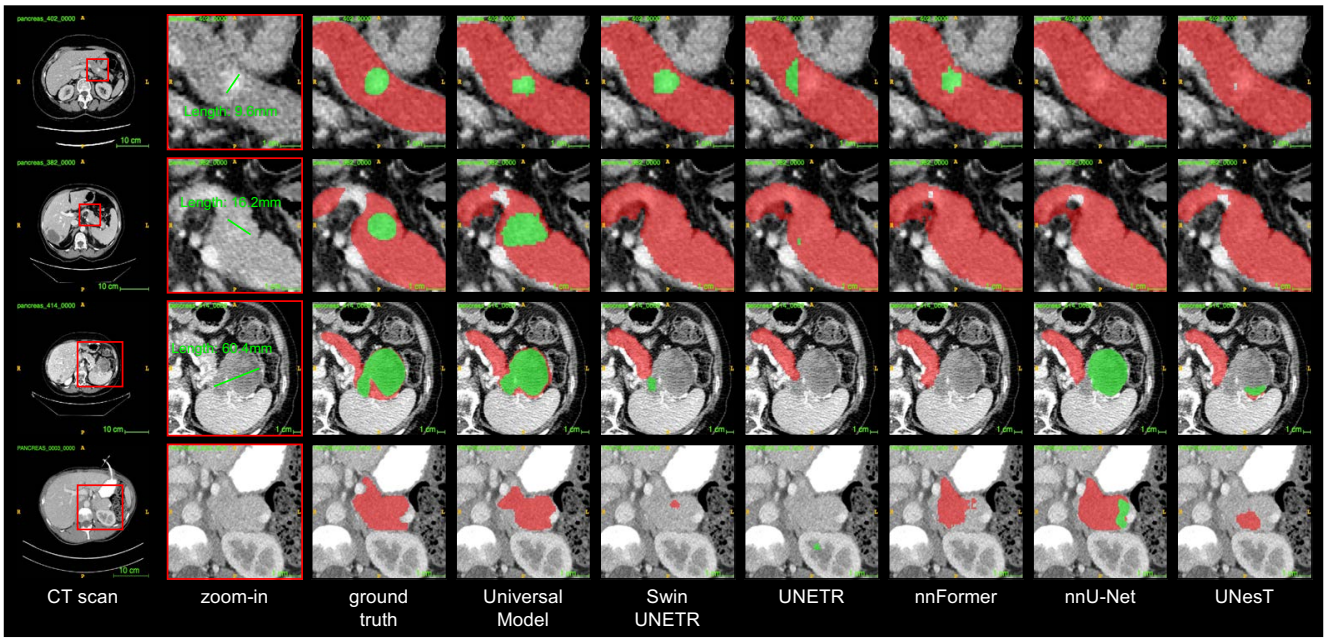


Figure 5. **Pancreatic tumor detection.** Qualitative visualizations of the proposed Universal Model and five competitive baseline methods. We review the detection results of tumors from smaller to larger sizes (Rows 1–3). When it comes to a CT scan without tumor from other hospitals, the Universal Model generalize well in organ segmentation and does not generate many false positives of tumors (Row 4; §4.2). The visualization of tumor detection in other organs (*e.g.*, liver tumors and kidney tumors) can be found in Appendix Figures 10–11.

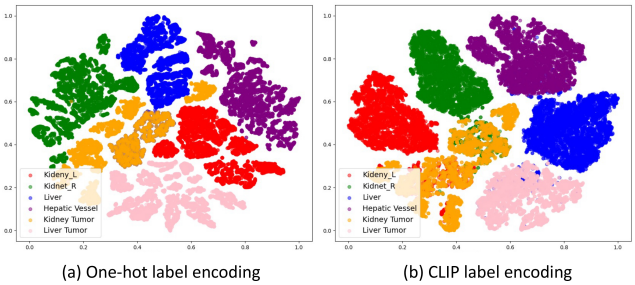
creatic, and colon tumors, leading to 14%, 8%, and 12% improvement in DSC scores, respectively. However, DSC scores cannot faithfully reveal the tumor detection performance because, by default, they are only calculated on abnormal CT scans (with tumors) [29]. The AI might generate numerous false positives when encountering normal CT scans (that have no tumor) [61]. Therefore, we also evaluate patient-level Sensitivity and Specificity for detecting the three types of tumors, and the harmonic mean of sensitivity and specificity is reported to indicate the balance between two abilities. To obtain normal CT scans, we adopt the CHAOS and Pancreas-CT datasets because these two datasets provide pathological verification that no tumors are present [71, 60]. Table 4 show that Universal Model achieves harmonic mean of 91.84%, 93.31% and 92.59% for three tumors, indicating the ability to accurately identify tumor cases while reducing false positives and achieving a competitive balance. Moreover, Rows 1–3 in Figure 5 depict the prediction of small/medium/large pancreatic tumors; Row 4 shows that Universal Model can precisely segment the pancreas and reduce the number of false positives on normal CT scans. 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.

4.3. Effectiveness of CLIP Embedding

We further show the t-SNE visualization of embedding space for both one-hot encoding and CLIP encoding in Figure 6. We can see that the decoder embedding of CLIP encoding shows better feature clustering and anatomical structure. For example, right kidney and left kidney features are closer in embedding space for universal model, which is ...ily matched with cosine similarity between CLIP em-

Table 4. **Tumor detection performance.** The CT scans in LiTS [3], KiTS [25], and MSD Pancreas [1] contain tumors in the liver, kidney, and pancreas, respectively. These scans are used to compute the sensitivity (Sen.) of tumor detection. To perform an alternative check of specificity (Spec.), we use CHAOS [71] and Pancreas-CT [60]. It has been confirmed that CHAOS has no liver or kidney tumor, and Pancreas-CT has no pancreatic tumor in the CT scans. The harmonic mean (Harm.) is calculated to indicate the balance between sensitivity and specificity. Universal Model achieves high harmonic mean, which is clinically important because it reveals that Universal Model can accurately identify tumor cases while reduce false positives.

Methods	Liver Tumor			Kidney Tumor			Pancreatic Tumor		
	Sen.	Spec.	Harm.	Sen.	Spec.	Harm.	Sen.	Spec.	Harm.
nnU-Net [29]	**94.44**	75.00	83.60	96.88	85.00	90.55	95.18	88.75	91.85
UNet++ [96]	**94.44**	80.00	86.62	N/A	N/A	N/A	N/A	N/A	N/A
UNETR [22]	86.11	**95.00**	90.34	93.75	**95.00**	**94.37**	90.36	81.25	85.56
Swin UNETR [68]	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**



(a) One-hot label encoding (b) CLIP label encoding

Figure 6. **t-SNE visualization of embedding space.** We compare the decoder embedding space of (a) One-hot label encoding and (b) CLIP label encoding with six categories, i.e., liver, liver tumor, right kidney, left kidney, kidney tumor and hepatic vessel, which is the same as in Figure 1. CLIP label encoding achieves a better feature cluster and shows anatomically structured semantics. Visualization of embedding space for all categories is provided in Appendix Figure 12.

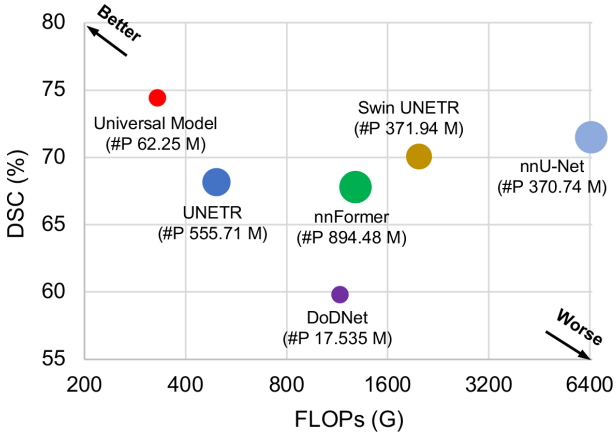


Figure 7. ***Efficiency:* FLOPs vs. DSC.** We plot the average DSC score on the 6 MSD tasks against the FLOPs (Floating-point operations per second). The FLOPs is computed based on input with spatial size $96 \times 96 \times 96$. The size of each circle indicates the number of parameters ('#P'). In the inference, Universal Model is faster than nnU-Net (2nd best in performance) and Swin UNETR (3rd best) by $19\times$ and $6\times$ measured by FLOPs, respectively.

beddings as shown in Figure 1. This validates that the CLIP-based encoding can facilitate the model to cap

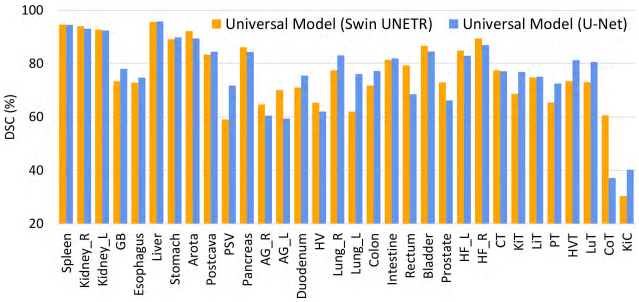


Figure 8. ***Expansibility:* flexible backbones.** Universal Model can be expanded to CNN-based (*e.g.*, U-Net [59]) and Transformer-based (*e.g.*, Swin UNETR [68]) backbone. For the abbreviation of some organs, please refer to Appendix Table 14. Both backbones achieve comparable results.

anatomical relationship and to learn a structured feature embedding. Furthermore, we conduct ablation study with various embedding to replace the CLIP embedding including BioLinkBERT embedding[8] [83], and the results are shown in Appendix Table 14. We can see that the CLIP-based embedding can significantly improve the performance comparing with conventional one-hot labels (DoDNet [88]) and text-only pre-trained embedding (BioLinkBERT [83]).

5. Intriguing Properties

Efficiency: FLOPs vs. DSC. It is clinically important to make AI models faster [9, 17]. The floating-point operations per second (FLOPS) are used to indicate the inference speed. Figure 7 presents a speed-performance plot, showing that Universal Model is computationally more efficient compared with dataset-specific models ($>6\times$ faster), while maintaining a high DSC score of 74% on average[9].

Expansibility: flexible backbones. The proposed Universal Model framework can be applied flexibly to other backbones. We further conduct experiments in CNN-based

[8] LinkBERT is a transformer encoder model pretrained on a large corpus of documents, which has capabilities for understanding medical text.

[9] Existing dataset-specific models are limited to being trained individu- for each MSD task, due to the partial label problem.

Table 5. ***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 dice score of the first seven organs. Compared with dataset-specific models, our Universal Model performs more robustly to CT scans taken from a variety of scanners, protocols, and institutes.

3D-IRCADb	spleen	kidneyR	kidneyL	gallbladder	liver	stomach	pancreas	lungR	lungL	mDSC*	mDSC
SegResNet [63]	94.08	80.01	91.60	69.59	95.62	**89.53**	79.19	N/A	N/A	85.66	N/A
nnFormer [91]	93.75	88.20	90.11	62.22	94.93	87.93	78.90	N/A	N/A	85.14	N/A
UNesT [87]	94.02	84.90	**94.95**	68.58	95.10	89.28	79.94	N/A	N/A	86.68	N/A
TransBTS [73]	91.33	76.22	88.87	62.50	94.42	85.87	63.90	N/A	N/A	80.44	N/A
TransUNet [7]	94.09	82.07	89.92	63.07	95.55	89.12	79.53	N/A	N/A	84.76	N/A
UNETR [22]	92.23	91.28	94.19	56.20	94.25	86.73	72.56	91.56	93.31	83.92	85.81
Swin UNETR [68]	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	gallbladder	liver	stomach	pancreas	arota	postcava	vein	mDSC
SegResNet [63]	93.11	89.92	87.84	74.62	95.37	87.90	76.33	84.05	79.36	57.13	82.56
nnFormer [91]	86.71	87.03	84.28	63.37	91.64	73.18	71.88	84.73	78.61	55.31	77.67
UNesT [87]	93.82	90.42	89.04	76.40	95.30	89.65	78.97	84.36	79.61	59.70	83.73
TransBTS [73]	85.47	81.58	82.00	60.58	92.50	72.29	63.25	83.47	75.07	55.38	75.16
TransUNet [7]	94.63	89.86	89.61	77.28	95.85	88.95	79.98	85.06	**81.02**	**59.76**	84.20
UNETR [22]	91.89	89.07	87.60	66.97	91.48	83.18	70.56	82.92	75.20	57.53	79.64
Swin UNETR [68]	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 6. ***Transferability:*** **Fine-tuning performance.** Fine-tuning Universal Model significantly outperforms learning from scratch on two downstream datasets (*i.e.*, TotalSegmentator and JHH). 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. Due to the space, the per-class evaluation of TotalSegmentator and JHH can be found in Appendix Tables 9–12 and Table 13, respectively.

Method	TotalSeg_vertebrae	TotalSeg_cardiac	TotalSeg_muscles	TotalSeg_organs	JHH_cardiac	JHH_organs
Scratch	81.06	84.47	88.83	86.42	71.63	89.08
MedicalNet [10]	82.28	87.40	91.36	86.90	58.07	77.68
Models Gen. [98]	85.12	86.51	89.96	85.78	**74.25**	88.64
Swin UNETR [68]	86.23	87.91	92.39	88.56	67.85	87.21
UniMiSS [77]	85.12	88.96	92.86	88.51	69.33	82.53
Universal Model	**86.49**	**89.57**	**94.43**	**88.95**	72.06	**89.37**

backbone (*i.e.*, U-Net [59]) and achieve an average DSC score of 76.73% over 25 organs and 6 tumors, which is comparable with the average DSC score of 76.11% obtained by Swin UNETR, as shown in Table 8.

Generalizability: results on external datasets. A key expectation of medical AI models is their generalizability, *i.e.*, performance on new data across many hospitals, rather than the performance tailored to a single dataset [46, 50, 26]. Compared with dataset-specific models, Universal Model has trained on the order of magnitude more diverse CT scans, therefore demonstrating significantly better generalizability. We conduct the evaluation on 3D-IRCADb (public) and JHH (private), which are absolutely not seen in the training. Universal Model substantially outperforms the previous methods on 3D-IRCADb and JHH with a DSC improvement of 5% and 4%, respectively (see Table 5).

Transferability: fine-tuning results. Universal Model can serve as a powerful pre-training model for segmentation. Through pre-training by assembly dataset directly and fine-tuning to other datasets, the Universal Model achieves the highest DSC compared with other pre-training methods with 86.49%, 89.57%, 94.43% and 88.95% for four downstream tasks in the TotalSegmentator dataset (see T

6. Conclusion

This work presents a CLIP-Driven Universal Model for abdominal organ segmentation and tumor detection. To address the label inconsistency and orthogonality problems, we integrate CLIP embedding with segmentation models, resulting in a flexible and powerful segmentor. The model can effectively learn from partially labeled datasets and achieve high performance, as evidenced by ranking first in both MSD and BTCV. The segmentation accuracy of six organs has approached that of humans. Importantly, our study demonstrates that CLIP embedding can establish a stronger and more meaningful anatomical relationship between organs and tumors than the widely-used one-hot embedding as the ground truth. Furthermore, we validate several clinically important merits of the CLIP-Driven Universal Model, including compelling efficiency, generalizability, transferability, and expansibility, through experimental results.

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, Tong Li, and Wenxuan Li for their constructive suggestions at several stages of the project.

References

[1] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, Bram van Ginneken, et al. The medical segmentation decathlon. *arXiv preprint arXiv:2106.05735*, 2021. 1, 7, 14, 15

[2] Xiaoyu Bai and Yong Xia. An end-to-end framework for universal lesion detection with missing annotations. In *2022 16th IEEE International Conference on Signal Processing (ICSP)*, volume 1, pages 411–415. IEEE, 2022. 2

[3] Patrick Bilic, Patrick Ferdinand Christ, Eugene Vorontsov, Grzegorz Chlebus, Hao Chen, Qi Dou, Chi-Wing Fu, Xiao Han, Pheng-Ann Heng, Jürgen Hesser, et al. The liver tumor segmentation benchmark (lits). *arXiv preprint arXiv:1901.04056*, 2019. 1, 7, 14, 15

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

[5] Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R Sabuncu, John Guttag, and Adrian V Dalca. Universeg: Universal medical image segmentation. *arXiv preprint arXiv:2304.06131*, 2023. 3

[6] Pierre Chambon, Christian Bluethgen, Curtis P Langlotz, and Akshay Chaudhari. Adapting pretrained vision-language foundational models to medical imaging domains. *arXiv preprint arXiv:2210.04133*, 2022. 2, 3

[7] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 3, 8

[8] Jieneng Chen, Yingda Xia, Jiawen Yao, Ke Yan, Jianpeng Zhang, Le Lu, Fakai Wang, Bo Zhou, Mingyan Qiu, Qihang Yu, et al. Towards a single unified model for effective detection, segmentation, and diagnosis of eight major cancers using a large collection of ct scans. *arXiv preprint arXiv:2301.12291*, 2023. 3

[9] Po-Hsuan Cameron Chen, Krishna Gadepalli, Robert MacDonald, Yun Liu, Shiro Kadowaki, Kunal Nagpal, Timo Kohlberger, Jeffrey Dean, Greg S Corrado, Jason D Hipp, et al. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nature medicine*, 25(9):1453–1457, 2019. 7

[10] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019. 8, 19, 20

[11] S Chen, K Ma, and Y Zheng. Transfer learning for 3d medical image analysis. *arXiv preprint arXiv*, 2019. 3

[12] Xuming Chen, Shanlin Sun, Narisu Bai, Kun Han, Qianqian Liu, Shengyu Yao, Hao Tang, Chupeng Zhang, Zhipeng Lu, Qian Huang, et al. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. *Radiotherapy and Oncology*, 160:175–184, 2021. 2

[13] Xuxin Chen, Ximin Wang, Ke Zhang, Kar-Ming Fung, Theresa C Thai, Kathleen Moore, Robert S Mannel, Hong Liu, Bin Zheng, and Yuchen Qiu. Recent advances and clinical applications of deep learning in medical image analysis. *Medical Image Analysis*, page 102444, 2022. 1

[14] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32, 2019. 3

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[16] Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906*, 2021. 3

[17] Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. *NPJ digital medicine*, 4(1):1–9, 2021. 7

[18] Xi Fang and Pingkun Yan. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. *IEEE Transactions on Medical Imaging*, 39(11):3619–3629, 2020. 2, 3

[19] Cheng Guo and Felix Berkhahn. Entity embeddings of categorical variables. *arXiv preprint arXiv:1604.06737*, 2016. 2

[20] Pengfei Guo, Puyang Wang, Jinyuan Zhou, Shanshan Jiang, and Vishal M Patel. 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*, pages 2423–2432, 2021. 4

[21] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B Gotway, and Jianming Liang. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE Transactions on Medical Imaging*, 2021. 5

[22] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022. 3, 6, 7, 8

[23] Yufan He, Dong Yang, Holger Roth, Can Zhao, and Daguang Xu. Dints: Differentiable neural network topology search for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5841–5850, 2021. 5

[24] Nicholas Heller, Sean McSweeney, Matthew Thomas Peterson, Sarah Peterson, Jack Rickman, Bethany Stai, Resha Tejpaul, Makinna Oestreich, Paul Blake, Joel Rosenberg, et al. An international challenge to use artificial intelligence to define the state-of-the-art in kidney and kidney tumor segmentation in ct imaging., 2020. 1, 14, 15

[25] Nicholas Heller, Niranjan Sathianathen, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel

Rosenberg, Paul Blake, Zachary Rengel, Makinna Oestreich, et al. The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445*, 2019. 7

[26] Qixin Hu, Yixiong Chen, Junfei Xiao, Shuwen Sun, Jieneng Chen, Alan L Yuille, and Zongwei Zhou. Label-free liver tumor segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7422–7432, 2023. 8

[27] Rui Huang, Yuanjie Zheng, Zhiqiang Hu, Shaoting Zhang, and Hongsheng Li. Multi-organ segmentation via co-training weight-averaged models from few-organ datasets. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 146–155. Springer, 2020. 3

[28] Ziyan Huang, Haoyu Wang, Zhongying Deng, Jin Ye, Yanzhou Su, Hui Sun, Junjun He, Yun Gu, Lixu Gu, Shaoting Zhang, et al. Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training. *arXiv preprint arXiv:2304.06716*, 2023. 3

[29] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. 2, 3, 4, 5, 6, 7

[30] Wei Ji, Shuang Yu, Junde Wu, Kai Ma, Cheng Bian, Qi Bi, Jingjing Li, Hanruo Liu, Li Cheng, and Yefeng Zheng. Learning calibrated medical image segmentation via multi-rater agreement modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12341–12351, 2021. 5

[31] Yuanfeng Ji, Haotian Bai, Jie Yang, Chongjian Ge, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2, 14, 15

[32] Mintong Kang, Bowen Li, Zengle Zhu, Yongyi Lu, Elliot K Fishman, Alan L Yuille, and Zongwei Zhou. Label-assemble: Leveraging multiple datasets with partial labels. In *IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2023. 1, 3

[33] Sungwoong Kim, Ildoo Kim, Sungbin Lim, Woonhyuk Baek, Chiheon Kim, Hyungjoo Cho, Boogeon Yoon, and Taesup Kim. Scalable neural architecture search for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 220–228. Springer, 2019. 5

[34] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 4

[35] Bennett Landman, Zhoubing Xu, Juan Eugenio Igelsias, Martin Styner, Thomas Robin Langerak, and Arno Klein. Multi-atlas labeling beyond the cranial vault-workshop and challenge. 2017. 2

[36] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, page 12, 2015. 14, 15

[37] Xin Li, Yiming Zhou, Zheng Pan, and Jiashi Feng. Partial order pruning: for best speed/accuracy trade-off in neural architecture search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9145–9153, 2019. 1

[38] Xiaokun Liang, Na Li, Zhicheng Zhang, Jing Xiong, Shoujun Zhou, and Yaoqin Xie. Incorporating the hybrid deformable model for improving the performance of abdominal ct segmentation via multi-scale feature fusion network. *Medical Image Analysis*, 73:102156, 2021. 2, 3

[39] Pengbo Liu, Yang Deng, Ce Wang, Yuan Hui, Qian Li, Jun Li, Shiwei Luo, Mengke Sun, Quan Quan, Shuxin Yang, et al. Universal segmentation of 33 anatomies. *arXiv preprint arXiv:2203.02098*, 2022. 2

[40] Zhe Liu, Kai Han, Kaifeng Xue, Yuqing Song, Lu Liu, Yangyang Tang, and Yan Zhu. Improving ct-image universal lesion detection with comprehensive data and feature enhancements. *Multimedia Systems*, pages 1–12, 2022. 2

[41] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096, 2022. 3

[42] Xiangde Luo, Wenjun Liao, Jianghong Xiao, Tao Song, Xiaofan Zhang, Kang Li, Guotai Wang, and Shaoting Zhang. Word: Revisiting organs segmentation in the whole abdominal region. *arXiv preprint arXiv:2111.02403*, 2021. 1, 2, 14, 15

[43] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. Abdomenct-1k: Is abdominal organ segmentation a solved problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 2, 14, 15

[44] Ruotian Ma, Xin Zhou, Tao Gui, Yiding Tan, Qi Zhang, and Xuanjing Huang. Template-free prompt tuning for few-shot ner. *arXiv preprint arXiv:2109.13532*, 2021. 14

[45] Tarun Mattikalli, Tejas Sudharshan Mathai, and Ronald M Summers. Universal lesion detection in ct scans using neural network ensembles. In *Medical Imaging 2022: Computer-Aided Diagnosis*, volume 12033, pages 864–868. SPIE, 2022. 2

[46] John Mongan, Linda Moy, and Charles E. Kahn. Checklist for artificial intelligence in medical imaging (claim): A guide for authors and reviewers. *Radiology: Artificial Intelligence*, 2(2):e200029, 2020. PMID: 33937821. 8

[47] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023. 3

[48] Varun Naga, Tejas Sudharshan Mathai, Angshuman Paul, and Ronald M Summers. Universal lesion detection and

classification using limited data and weakly-supervised self-training. In *Workshop on Medical Image Learning with Limited and Noisy Data*, pages 55–64. Springer, 2022. 2

[49] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernardino Romera-Paredes, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv preprint arXiv:1809.04430*, 2018. 16

[50] Beau Norgeot, Giorgio Quer, Brett K Beaulieu-Jones, Ali Torkamani, Raquel Dias, Milena Gianfrancesco, Rima Arnaout, Isaac S Kohane, Suchi Saria, Eric Topol, et al. Minimum information about clinical artificial intelligence modeling: the mi-claim checklist. *Nature medicine*, 26(9):1320–1324, 2020. 8

[51] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018. 2

[52] Mauricio Orbes-Arteaga, Thomas Varsavsky, Carole H Sudre, Zach Eaton-Rosen, Lewis J Haddow, Lauge Sørensen, Mads Nielsen, Akshay Pai, Sébastien Ourselin, Marc Modat, et al. 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*, pages 54–62. Springer, 2019. 4

[53] Kwanyong Park, Sanghyun Woo, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. Per-clip video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1352–1361, 2022. 3

[54] Francesco Piccialli, Vittorio Di Somma, Fabio Giampaolo, Salvatore Cuomo, and Giancarlo Fortino. A survey on deep learning in medicine: Why, how and when? *Information Fusion*, 66:111–137, 2021. 1

[55] Ziyuan Qin, Huahui Yi, Qicheng Lao, and Kang Li. Medical image understanding with pretrained vision language models: A comprehensive study. *arXiv preprint arXiv:2209.15517*, 2022. 3

[56] Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Jie Liu, Yucheng Tang, Alan Yuille, and Zongwei Zhou. Annotating 8,000 abdominal ct volumes for multi-organ segmentation in three weeks. *arXiv preprint arXiv:2305.09666*, 2023. 3

[57] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 3

[58] Blaine Rister, Darvin Yi, Kaushik Shivakumar, Tomomi Nobashi, and Daniel L Rubin. Ct-org, a new dataset for multiple organ segmentation in computed tomography. *Scientific Data*, 7(1):1–9, 2020. 14, 15

[59] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 2, 3, 7, 8

[60] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 556–564. Springer, 2015. 6, 7, 14, 15

[61] Yiqiu Shen, Farah E Shamout, Jamie R Oliver, Jan Witowski, Kawshik Kannan, Jungkyu Park, Nan Wu, Connor Huddleston, Stacey Wolfson, Alexandra Millet, et al. Artificial intelligence system reduces false-positive findings in the interpretation of breast ultrasound exams. *Nature communications*, 12(1):1–13, 2021. 6

[62] Gonglei Shi, Li Xiao, Yang Chen, and S Kevin Zhou. Marginal loss and exclusion loss for partially supervised multi-organ segmentation. *Medical Image Analysis*, 70:101979, 2021. 2, 3

[63] Md Mahfuzur Rahman Siddiquee and Andriy Myronenko. Redundancy reduction in semantic segmentation of 3d brain tumor mris. *arXiv preprint arXiv:2111.00742*, 2021. 8

[64] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022. 3

[65] L Soler, A Hostettler, V Agnus, A Charnoz, J Fasquel, J Moreau, A Osswald, M Bouhadjar, and J Marescaux. 3d image reconstruction for comparison of algorithm database: A patient specific anatomical and medical image database. *IRCAD, Strasbourg, France, Tech. Rep*, 2010. 15

[66] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey N Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Medical Image Analysis*, page 101693, 2020. 1

[67] Yucheng Tang, Riqiang Gao, Ho Hin Lee, Shizhong Han, Yunqiang Chen, Dashan Gao, Vishwesh Nath, Camilo Bermudez, Michael R Savona, Richard G Abramson, et al. High-resolution 3d abdominal segmentation with random patch network fusion. *Medical image analysis*, 69:101894, 2021. 6

[68] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. 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*, pages 20730–20740, 2022. 3, 5, 6, 7, 8, 15, 16, 19, 20

[69] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *European conference on computer vision*, pages 282–298. Springer, 2020. 4

[70] Constantin Ulrich, Fabian Isensee, Tassilo Wald, Maximilian Zenk, Michael Baumgartner, and Klaus H Maier-Hein.

Multitalent: A multi-dataset approach to medical image segmentation. *arXiv preprint arXiv:2303.14444*, 2023. 3

[71] Vanya V Valindria, Nick Pawlowski, Martin Rajchl, Ioannis Lavdas, Eric O Aboagye, Andrea G Rockall, Daniel Rueckert, and Ben Glocker. 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)*, pages 547–556. IEEE, 2018. 6, 7, 14, 15

[72] Shanshan Wang, Cheng Li, Rongpin Wang, Zaiyi Liu, Meiyun Wang, Hongna Tan, Yaping Wu, Xinfeng Liu, Hui Sun, Rui Yang, et al. Annotation-efficient deep learning for automatic medical image segmentation. *Nature communications*, 12(1):1–13, 2021. 1

[73] Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 109–119. Springer, 2021. 8

[74] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11686–11695, 2022. 3

[75] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022. 3

[76] Jakob Wasserthal, Manfred Meyer, Hanns-Christian Breit, Joshy Cyriac, Shan Yang, and Martin Segeroth. Totalsegmentator: robust segmentation of 104 anatomical structures in ct images. *arXiv preprint arXiv:2208.05868*, 2022. 1, 2, 15

[77] Yutong Xie, Jianpeng Zhang, Yong Xia, and Qi Wu. Unimiss: Universal medical self-supervised learning via breaking dimensionality barrier. In *European Conference on Computer Vision*, pages 558–575. Springer, 2022. 8, 19, 20

[78] Ke Yan, Jinzheng Cai, Adam P Harrison, Dakai Jin, Jing Xiao, and Le Lu. Universal lesion detection by learning from multiple heterogeneously labeled datasets. *arXiv preprint arXiv:2005.13753*, 2020. 2

[79] Ke Yan, Jinzheng Cai, Youjing Zheng, Adam P Harrison, Dakai Jin, You-bao Tang, Yu-Xing Tang, Lingyun Huang, Jing Xiao, and Le Lu. Learning from multiple datasets with heterogeneous and partial labels for universal lesion detection in ct. *IEEE Transactions on Medical Imaging*, 2020. 3

[80] Ke Yan, Youbao Tang, Yifan Peng, Veit Sandfort, Mohammadhadi Bagheri, Zhiyong Lu, and Ronald M Summers. Mulan: multitask universal lesion analysis network for joint lesion detection, tagging, and segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 194–202. Springer, 2019. 2

[81] Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging*, 5(3):036501, 2018. 3

[82] Wenjun Yan, Lu Huang, Liming Xia, Shengjia Gu, Fuhua Yan, Yuanyuan Wang, and Qian Tao. Mri manufacturer shift and adaptation: increasing the generalizability of deep learning segmentation for mr images acquired with different scanners. *Radiology: Artificial Intelligence*, 2(4):e190195, 2020. 4

[83] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827*, 2022. 4, 7, 21

[84] Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, and Yong Xia. Uniseg: A prompt-driven universal segmentation model as well as a strong representation learner. *arXiv preprint arXiv:2304.03493*, 2023. 3

[85] Huahui Yi, Ziyuan Qin, Qicheng Lao, Wei Xu, Zekun Jiang, Dequan Wang, Shaoting Zhang, and Kang Li. Towards general purpose medical ai: Continual learning medical foundation model. *arXiv preprint arXiv:2303.06580*, 2023. 3

[86] Qihang Yu, Dong Yang, Holger Roth, Yutong Bai, Yixiao Zhang, Alan L Yuille, and Daguang Xu. 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*, pages 4126–4135, 2020. 5

[87] Xin Yu, Qi Yang, Yinchi Zhou, Leon Y Cai, Riqiang Gao, Ho Hin Lee, Thomas Li, Shunxing Bao, Zhoubing Xu, Thomas A Lasko, et al. Unest: Local spatial representation learning with hierarchical transformer for efficient medical segmentation. *arXiv preprint arXiv:2209.14378*, 2022. 8

[88] Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1195–1204, 2021. 1, 2, 3, 4, 7, 21

[89] Wenhua Zhang, Jun Zhang, Xiyue Wang, Sen Yang, Junzhou Huang, Wei Yang, Wenping Wang, and Xiao Han. Merging nucleus datasets by correlation-based cross-training. *Medical Image Analysis*, page 102705, 2022. 2

[90] Yixiao Zhang, Xinyi Li, Huimiao Chen, Alan Yuille, Yaoyao Liu, and Zongwei Zhou. Continual learning for abdominal multi-organ and tumor segmentation. *International conference on medical image computing and computer-assisted intervention*, 2023. 3

[91] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021. 3, 6, 8

[92] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 14

[93] Yuyin Zhou, Zhe Li, Song Bai, Chong Wang, Xinlei Chen, Mei Han, Elliot Fishman, and Alan L Yuille. Prior-aware neural network for partially-supervised multi-organ segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10672–10681, 2019. 2, 3

[94] Zongwei Zhou. *Towards Annotation-Efficient Deep Learning for Computer-Aided Diagnosis*. PhD thesis, Arizona State University, 2021. 1

[95] Zongwei Zhou, Michael B Gotway, and Jianming Liang. Interpreting medical images. In *Intelligent Systems in Medicine and Health*, pages 343–371. Springer, 2022. 1

[96] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6):1856–1867, 2019. 2, 3, 7

[97] Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021. 2, 5

[98] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3d medical image analysis. In *International conference on medical image computing and computer-assisted intervention*, pages 384–393. Springer, 2019. 8, 19, 20

[99] Zengle Zhu, Mintong Kang, Alan Yuille, and Zongwei Zhou. Assembling and exploiting large-scale existing labels of common thorax diseases for improved covid-19 classification using chest radiographs. In *Radiological Society of North America (RSNA)*, 2022. 3

[100] Martin Zlocha, Ben Glocker, and Jonathan Passerat-Palmbach. Universal lesion detector: Deep learning for analysing medical scans. 2019. 2