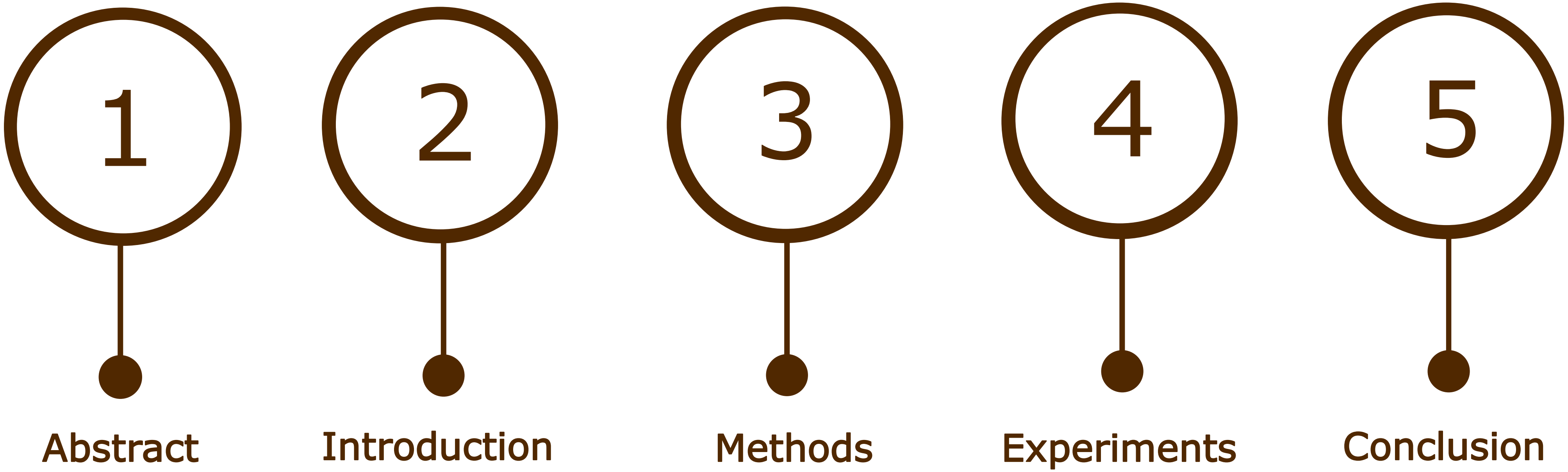
CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection

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

Presenter: Sunggu Kyung

Email: babbu3682@gmail.com

Paper Contents



1 Abstract

2 Introduction

3 Methods

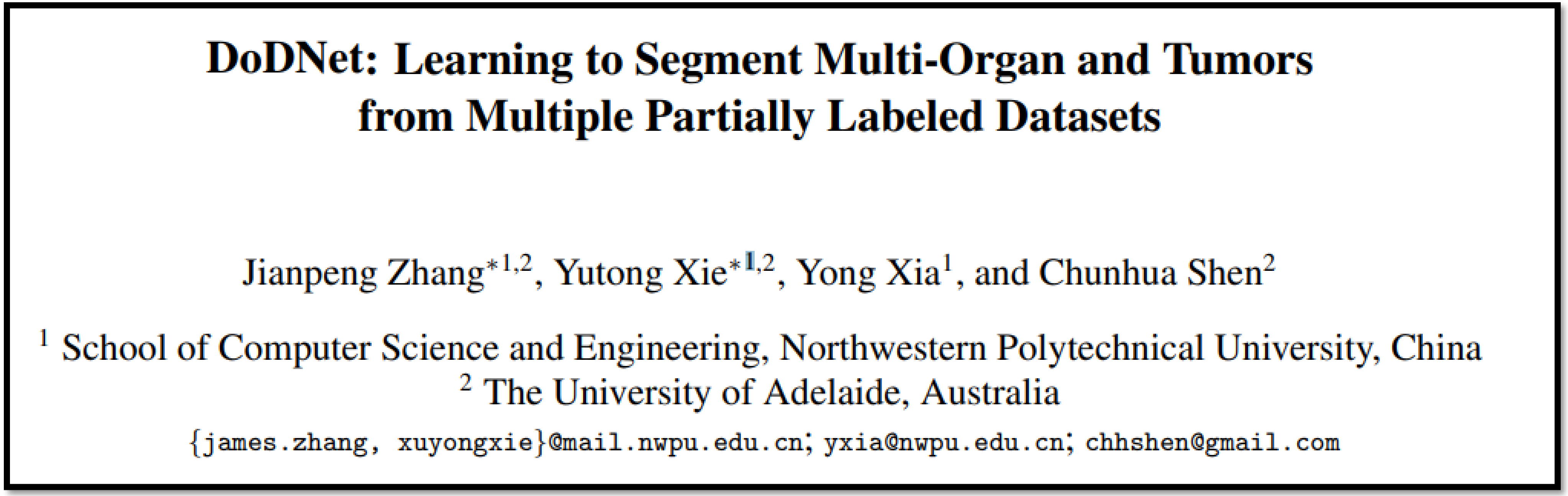
4 Experiments

5 Conclusion

Background

DoDNet

DoDNet (CVPR2021)



DoDNet: Learning to Segment Multi-Organ and Tumors from Multiple Partially Labeled Datasets

Jianpeng Zhang*[1,2], Yutong Xie*[1,2], Yong Xia[1], and Chunhua Shen[2]

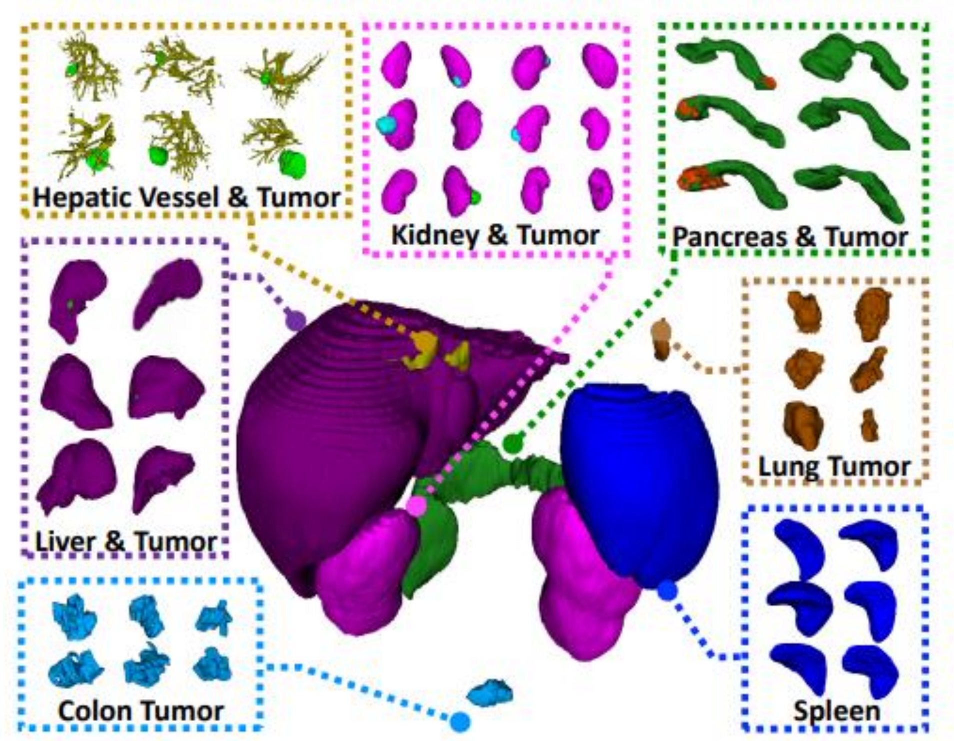
[1] School of Computer Science and Engineering, Northwestern Polytechnical University, China
[2] The University of Adelaide, Australia

{james.zhang, xuyongxie}@mail.nwpu.edu.cn; yxia@nwpu.edu.cn; chhshen@gmail.com

# Git: https://github.com/jianpengz/DoDNet

DoDNet (CVPR2021)

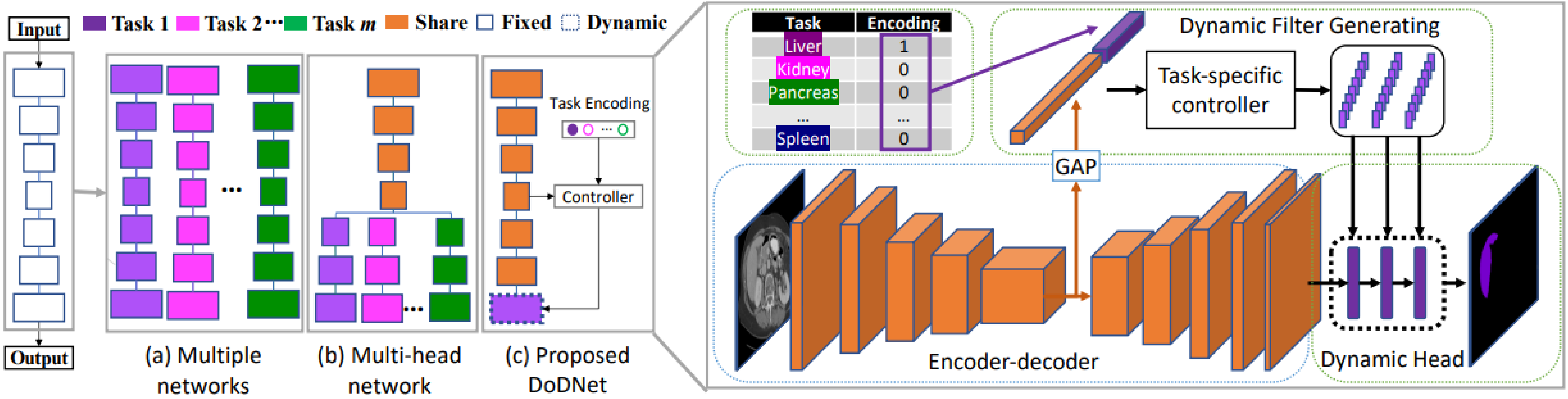
- partially labeled multi-organ and tumor segmentation



# DoDNet: Learning to Segment Multi-Organ and Tumors from Multiple Partially Labeled Datasets

DoDNet (CVPR2021)

- Overview



Task	Encoding
Liver	1
Kidney	0
Pancreas	0
...	...
Spleen	0

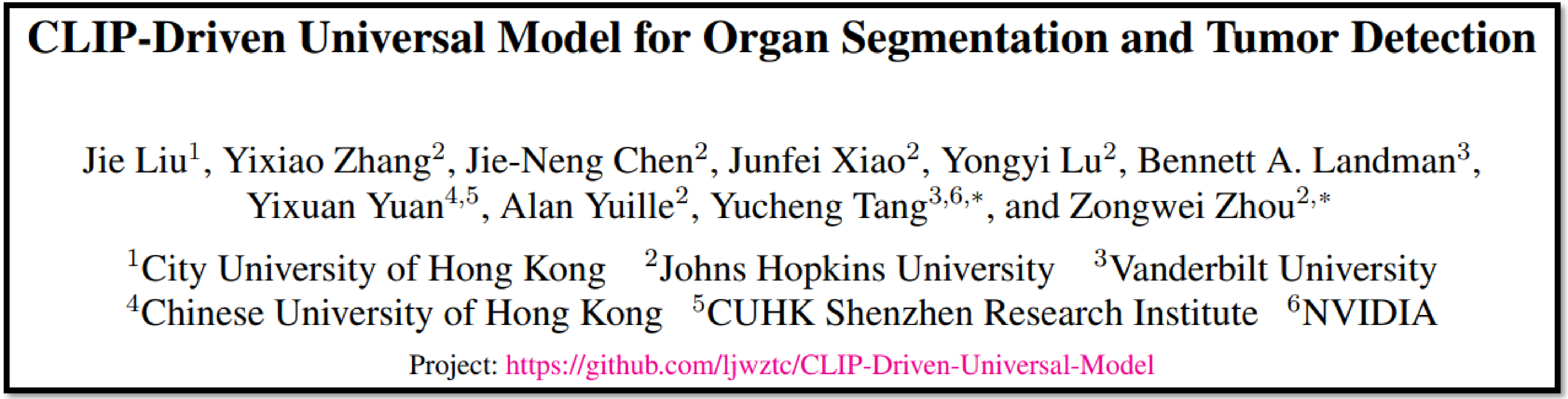
(a) Multiple networks

(b) Multi-head network

(c) Proposed DoDNet

Figure 2 – Three types of methods to perform m partially labeled segmentation tasks. (a) Multiple networks: Training m networks on m partially labeled subsets, respectively; (b) Multi-head networks: Training one network that consists of a shared encoder and m task-specific decoders (heads), each performing a partially labeled segmentation task; and (c) Proposed DoDNet: It has an encoder, a task encoding module, a dynamic filter generation module, and a dynamic segmentation head. The kernels in the dynamic head are conditioned on the input image and assigned task.

# DoDNet: Learning to Segment Multi-Organ and Tumors from Multiple Partially Labeled Datasets



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

- Due to the **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.

- 의료 영상 분야는 점점 늘어나는 annotated 데이터셋 덕분에 상당한 발전을 이루었지만, 세부적인 annotate 작성 비용이 높아 Robust한 AI 모델을 개발하는 데 있어서는 여전히 데이터셋이 작다는 인식이 있습니다.

- 부분적으로 라벨링된 데이터셋에 따른 제한이 다기관 세분화 및 종양 탐지를 위한 모델 성능에 큰 장애를 주고 있습니다.

- 이런 어려움에도 불구하고, 이런 분야에서 AI 모델의 잠재력은 높으며, 아직까지는 대부분 탐구 되지 않았습니다.

- 14개의 공개 데이터셋을 조합하여, 모델 확장성, 일반화, 그리고 전이성을 중점으로 한 AI 프레임워크의 임상적 영향을 보여주고자 했습니다.

- 부분적으로 주석이 달린 데이터셋을 조합하는 것에는 다섯 가지 측면에서의 라벨 불일치, 즉, 인덱스 불일치, 이름 불일치, 배경 불일치, 기관 중첩, 그리고 데이터 중첩 등의 문제가 있습니다.

- 라벨 직교성 문제도 있으며, 현재 세분화 방법들은 one-hot 라벨을 사용하여 클래스 간의 의미적 관계를 무시하고 있습니다.

- few-hot label은 가능한 해결책을 제공할 수 있으며, 이를 통해 간암이 간의 일부임을 나타낼 수 있지만, 기

Ours...

- 의료 영상 분야에서의 도전 과제를 해결하기 위해, 텍스트 임베딩을 통합하고 이진 분할 마스크를 사용하는 masked back-propagation 메커니즘을 채택하는 CLIP 기반 유니버설 모델이 제안되었습니다.

- 아키텍처는 Guo 등의 연구에서 영감을 받아, one-hot label 또는 few-hot label을 CLIP의 사전 훈련된 텍스트 인코더로 생성된 텍스트 임베딩으로 대체합니다.

- 이 CLIP 기반 라벨 인코딩은 유니버설 모델 피처 임베딩의 해부학적 구조를 강화합니다.

- 손실 계산은 사용 가능한 클래스에 대해서만 수행됩니다.

- 제안된 CLIP 기반 유니버설 모델은 최첨단 성능으로 25개의 기관 분할과 6개의 종양 탐지를 우수하게 수행할 수 있으며, 다른 기관에서의 CT 스캔에도 일반화될 수 있습니다.

Ours...

- 이 모델은 여섯 가지 장점을 가지고 있습니다:

1. 높은 복부 기관 세분화 성능
2. 높은 민감도를 유지하면서 적은 양의 거짓 긍정 예측
3. 계산 효율성이 높아, 테스트 속도를 6배 가속화
4. 다양한 백본으로 모델 확장 가능
5. 추가 조정과 적응 없이 다양한 병원에서의 CT 스캔에 대한 기관 세분화와 종양 탐지 성능의 일반화;
6. 다양한 하류 작업에 대한 효과적인 기반 모델로, 여러 질병, 기관, 데이터셋을 걸쳐 강력한 전이성

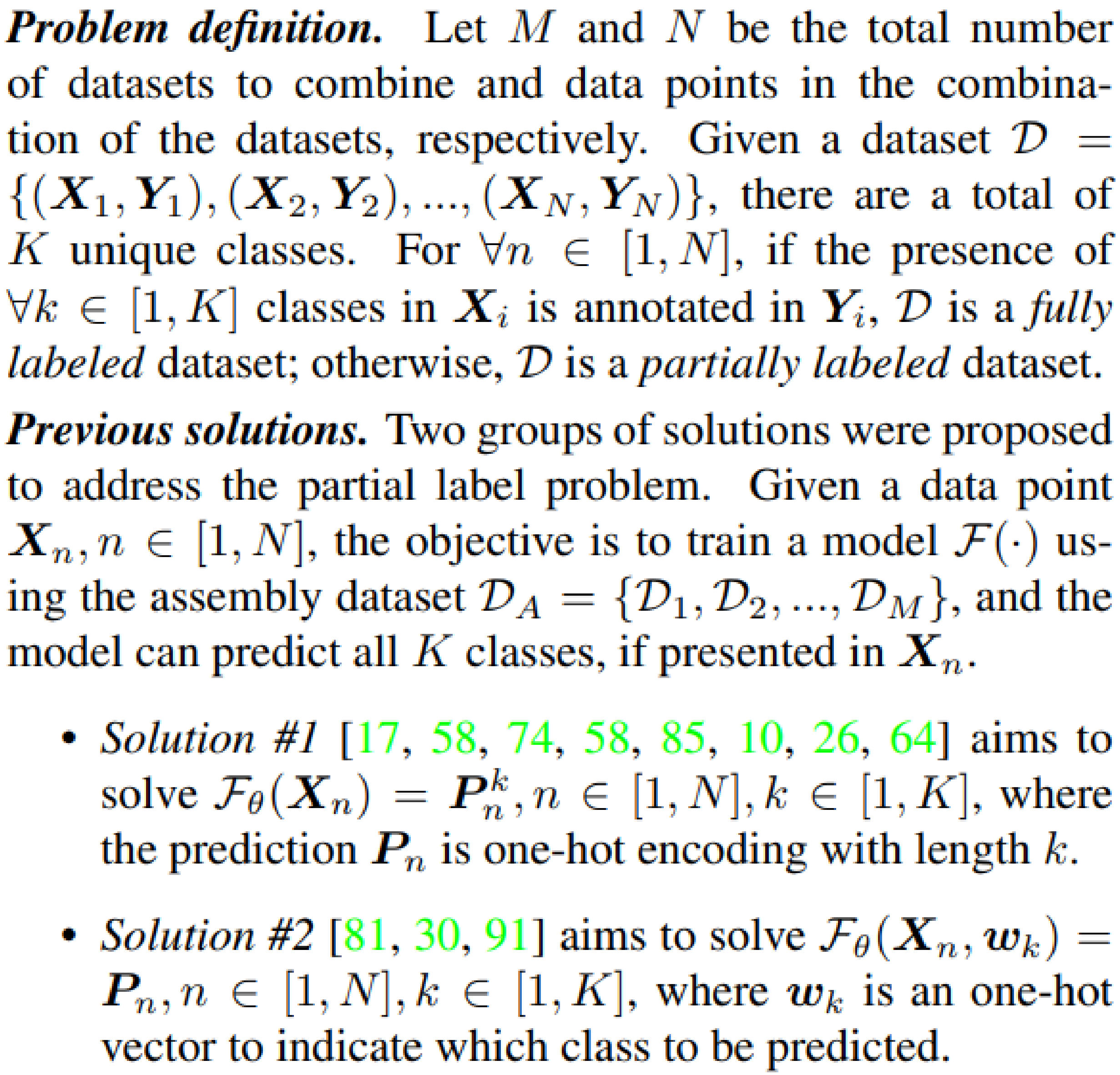
Related Work...

Partial label problem.
공개적으로 이용 가능한 복부 영상 데이터셋은 다른 기관 및 종양에 초점을 맞추고 있으며, 이들의 라벨 체계가 일관되지 않기 때문에 이러한 데이터셋의 조합을 통해 AI 모델을 학습시킬 때 부분 라벨 문제가 발생합니다. 이러한 한계를 극복하기 위해, CLIP 임베딩의 도입이 우리가 제안하는 프레임워크에서 중요한 요소임이 확인되었습니다.

Organ segmentation and tumor detection.
딥러닝 기반의 방법들은 기관 세분화와 종양 탐지에 널리 적용되어 왔습니다. 이와 다르게 Universal Model은 CLIP 임베딩을 도입하여 기관과 종양 간의 의미적 관계를 포착하며, 한 가지 프레임워크 내에서 여러가지 task들을 동시에 수행합니다.

CLIP in medical imaging.
언어 처리 및 이해 분야에서 대형 모델의 성공을 바탕으로, 최근에는 시각-언어 모델이 다양한 시각적 작업에 적용되었지만, 의료 분야에는 드물게 적용되었습니다. 우리는 이러한 방법을 이용하여 CLIP 임베딩을 의료 작업, 즉, 세분화에 도입하여 해부학적 구조 간의 의미적 관계의 중요성을 강조하고 있습니다.

Background.

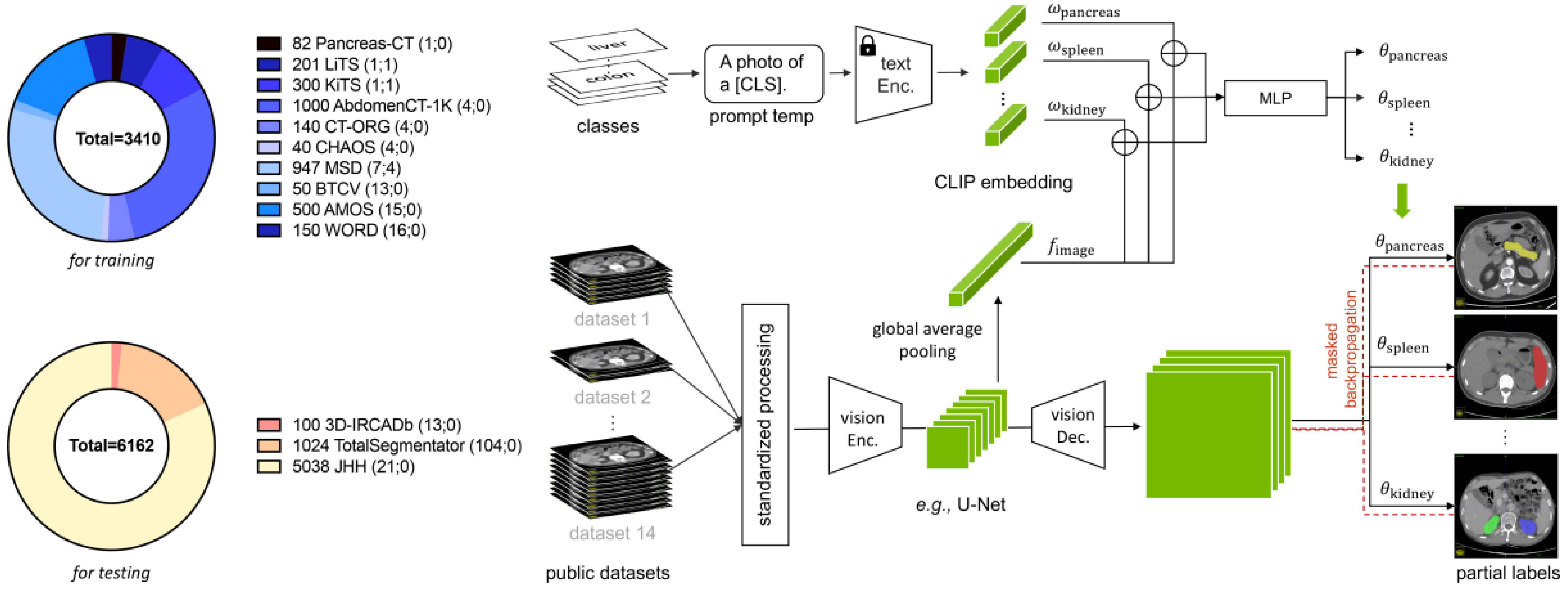


Problem definition. Let M and N be the total number of datasets to combine and data points in the combination of the datasets, respectively. Given a dataset $\mathcal{D} = \{(\boldsymbol{X}_1, \boldsymbol{Y}_1), (\boldsymbol{X}_2, \boldsymbol{Y}_2), ..., (\boldsymbol{X}_N, \boldsymbol{Y}_N)\}$, there 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.

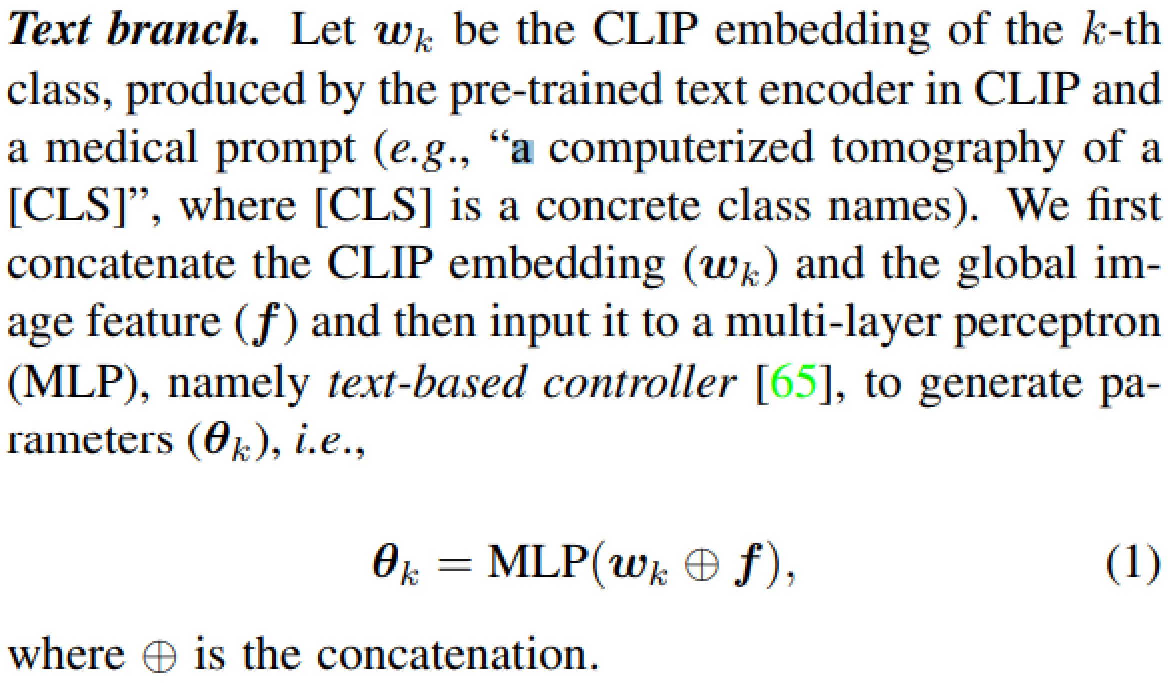
Previous solutions. Two groups of solutions were proposed to address the partial label problem. Given a data point $\boldsymbol{X}_n, n \in [1, N]$, the objective is to train a model $\mathcal{F}(\cdot)$ using the assembly dataset $\mathcal{D}_A = \{\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_M\}$,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

- *Solution #1* [17, 58, 74, 58, 85, 10, 26, 64] 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* [81, 30, 91] 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.

CLIP-Driven Universal Model - Architecture



- Text branch



Text branch. Let 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 names). We first concatenate the CLIP embedding (w_k) and the global image feature (f) and then input it to a multi-layer perceptron (MLP), namely *text-based controller* [65], to generate parameters (θ_k), *i.e.*,

$$\theta_k = \mathrm{MLP}(w_k \oplus f), \tag{1}$$

where $\oplus$ is the concatenation.

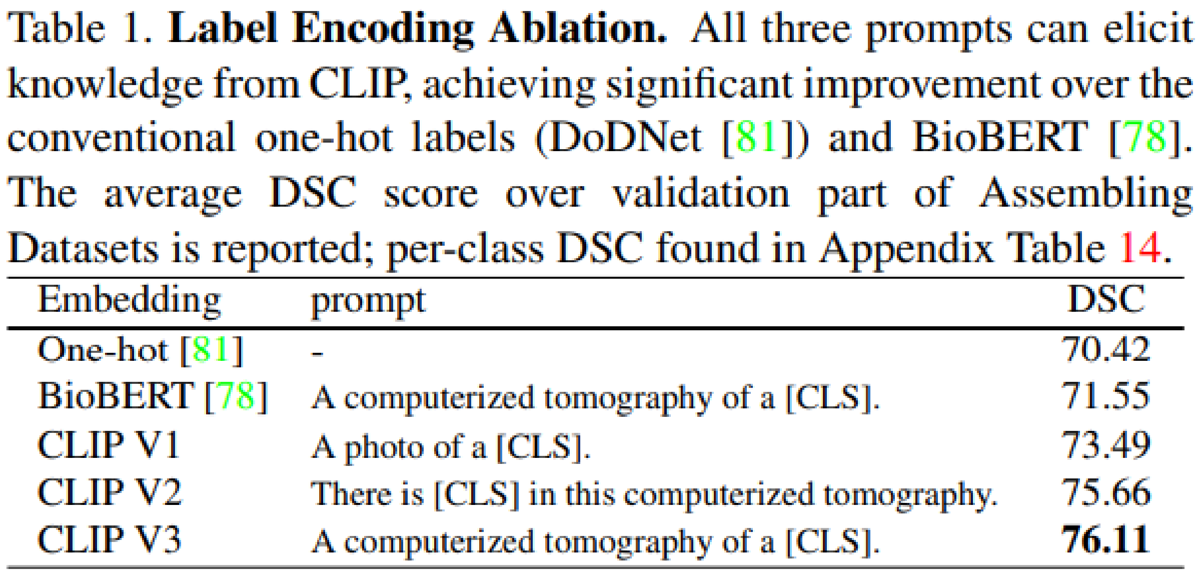


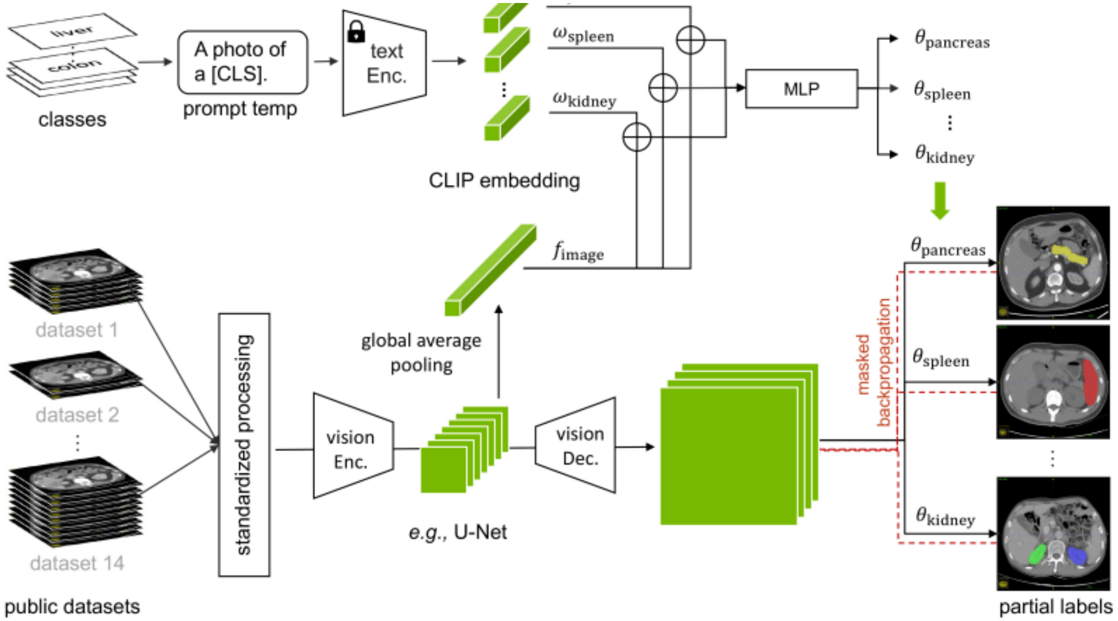
Table 1. **Label Encoding Ablation.** All three prompts can elicit knowledge from CLIP, achieving significant improvement over the conventional one-hot labels (DoDNet [81]) and BioBERT [78]. The average DSC score over validation part of Assembling Datasets is reported; per-class DSC found in Appendix Table 14.

Embedding	prompt	DSC
One-hot [81]	-	70.42
BioBERT [78]	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**

- **Vision branch**

$$\boldsymbol{P}_k = \mathrm{Sigmoid}\left(\left(\left(\boldsymbol{F} * \boldsymbol{\theta}_{k_1}\right) * \boldsymbol{\theta}_{k_2}\right) * \boldsymbol{\theta}_{k_3}\right), \qquad (2)$$

where $\boldsymbol{\theta}_{k_1}, \boldsymbol{\theta}_{k_2}, \boldsymbol{\theta}_{k_3}$ are computed by Equation 1, and $*$ represents the convolution. For each class $[\mathrm{CLS}]_k$, we generate the prediction using *one vs. all* manner (*i.e.*, Sigmoid instead of Softmax).



- **Masked back-propagation**

- 라벨 불일치 문제를 해결하기 위해, 우리는 마스크된 역전파 기법을 제안합니다.
- BCE 손실 함수가 감독에 사용됩니다.
- Y에 포함되지 않은 클래스의 손실 항을 마스킹하고 정확한 감독을 업데이트하기 위해 전체 프레임워크를 역전파합니다.
- 마스크된 역전파는 부분 라벨 문제에서 라벨 불일치를 해결합니다.
- 특히, 부분적으로 라벨링된 데이터셋은 다른 기관들을 배경으로 표시하며, 이는 기존 훈련 (Solution #1)의 능력을 제한합니다.

# https://github.com/ljwztc/CLIP-Driven-Universal-Model/blob/main/utils/loss.py

Experiments

- Quantitative Evaluation

  - ‘Vicunan’에서 제시한 방법을 바탕으로, 우리는 GPT-4를 활용하여 우리 모델의 생성된 응답의 품질을 측정.

  - 구체적으로, 우리는 COCO 검증 분할에서 무작위로 30개의 이미지를 선택하고, 제안된 데이터 생성 파이프라인을 사용하여 세 가지 타입의 질문(대화, 상세 설명, 복잡한 추론)을 생성합니다.

  - GPT-4는 [question, ground-truth bounding box, caption]을 바탕으로 reference prediction을 만들어, teacher 모델 역할로 상한선을 확보합니다. 두 모델에서의 응답을 얻은 후, 우리는 질문, 시각 정보 (캡션과 경계 상자의 형식), 그리고 두 assistant로부터 생성된 응답들을 GPT-4에 제공합니다.

  - GPT-4는 assistant로부터 받은 응답의 helpfulness, relevance, accuracy, level of details을 평가하고, 1에서 10까지의 척도에서 전반적인 점수를 주며, 높은 점수는 더 나은 성능을 나타냅니다.

Experiments

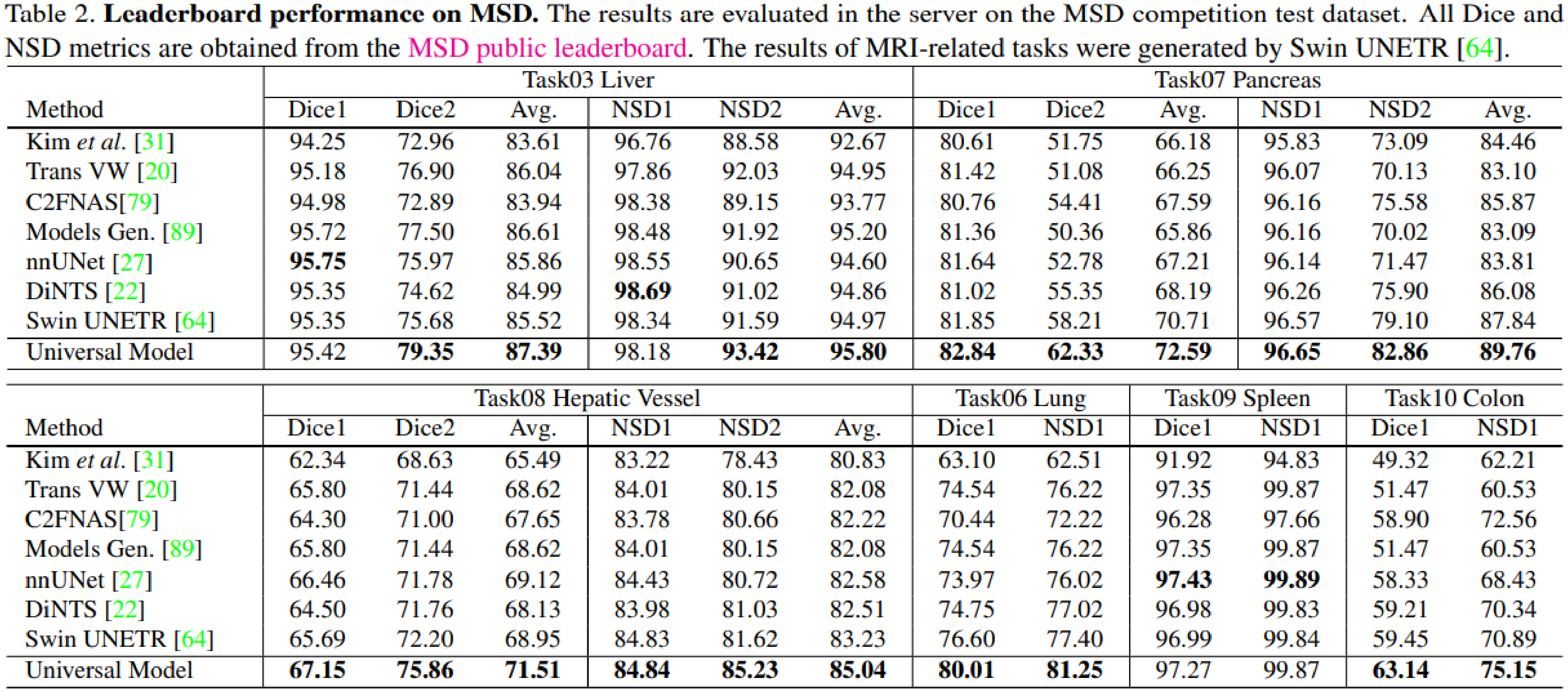
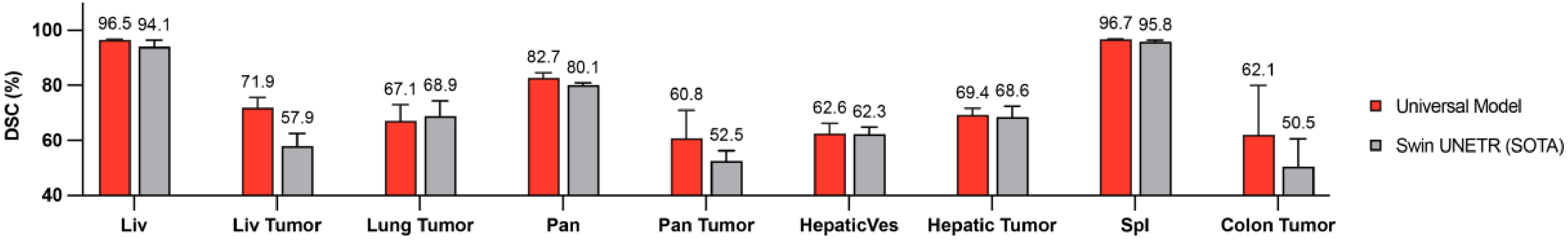


Table 2. **Leaderboard performance on MSD.** The results are evaluated in the server on the MSD competition test dataset. All Dice and NSD metrics are obtained from the MSD public leaderboard. The results of MRI-related tasks were generated by Swin UNETR [64].

Method	Task03 Liver Dice1	Dice2	Avg.	NSD1	NSD2	Avg.	Task07 Pancreas Dice1	Dice2	Avg.	NSD1	NSD2	Avg.
Kim *et al.* [31]	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 [20]	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[79]	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. [89]	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 [27]	**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 [22]	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 [64]	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 Dice1	Dice2	Avg.	NSD1	NSD2	Avg.	Task06 Lung Dice1	NSD1	Task09 Spleen Dice1	NSD1	Task10 Colon Dice1	NSD1
Kim *et al.* [31]	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 [20]	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[79]	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. [89]	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 [27]	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 [22]	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 [64]	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**



Experiments

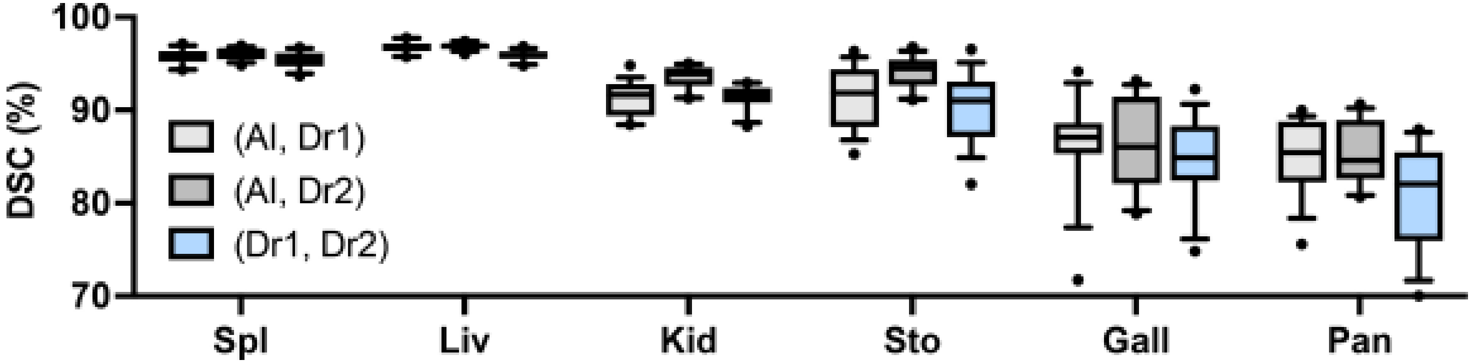


Figure 4. **Intra-observer variability.** We obtain similar performance between pseudo labels generated by the Universal Model (AI) and annotations performed by two human experts (Dr1,2) on 6 organs. Spleen (Spl), liver (Liv), kidneys (Kid), stomach (Sto), gallbladder (Gall), and pancreas (Pan) can be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

Experiments

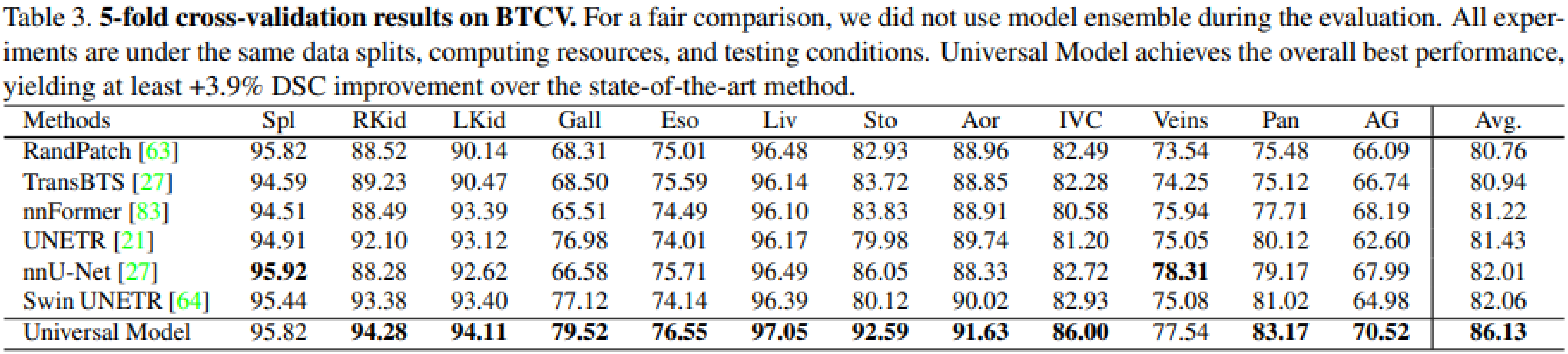


Table 3. **5-fold cross-validation results on BTCV.** For a fair comparison, we did not use model ensemble during the evaluation. All experiments are under the same data splits, computing resources, and testing conditions. Universal Model achieves the overall best performance, yielding at least +3.9%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

Methods	Spl	RKid	LKid	Gall	Eso	Liv	Sto	Aor	IVC	Veins	Pan	AG	Avg.
RandPatch [63]	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 [27]	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 [83]	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 [21]	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 [27]	**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 [64]	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**

Experiments

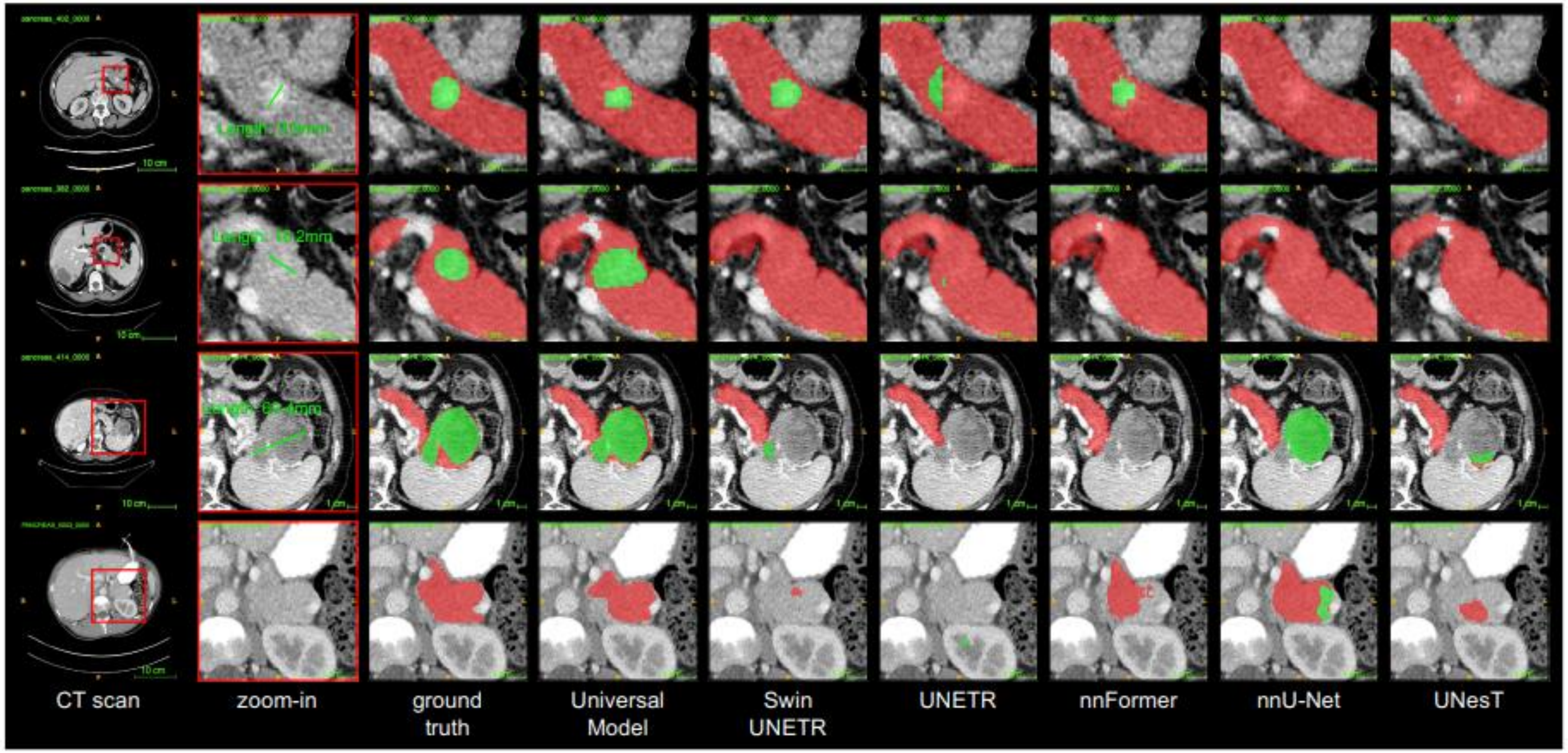


Figure 5. **Pancreatic tumor detection.** Qualitative visualizations of the proposed Universal Model and five competitive baseline methods. We review the detection results of tumors from smaller to larger sizes (Rows 1–3). When it comes a CT scan without tumor from other hospitals, the Universal Model 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.

Experiments

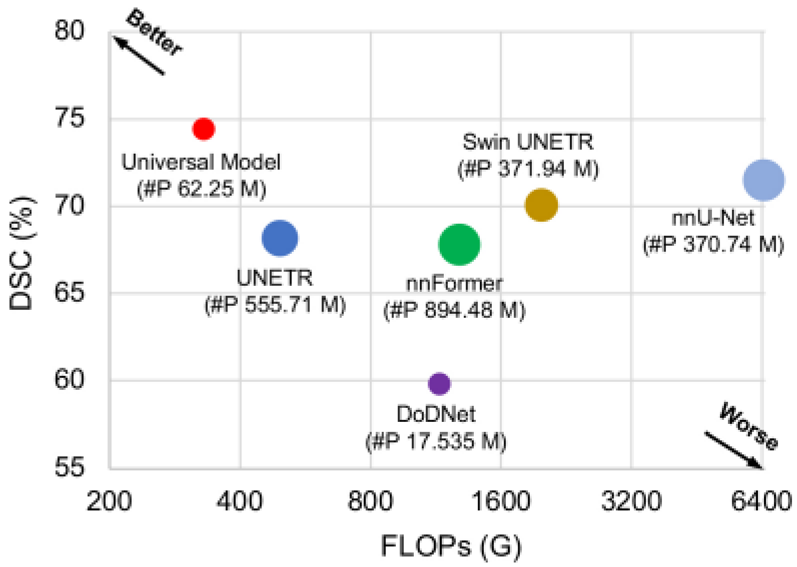


Figure 7. ***Efficiency:* FLOPs vs. DSC.** We plot the average DSC score on the 6 MSD tasks against the FLOPs (Floating-point operations per second). The FLOPs is computed based on input with spatial size 96 × 96 × 96. The size of each circle indicates the number of parameters ('#P'). In the inference,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

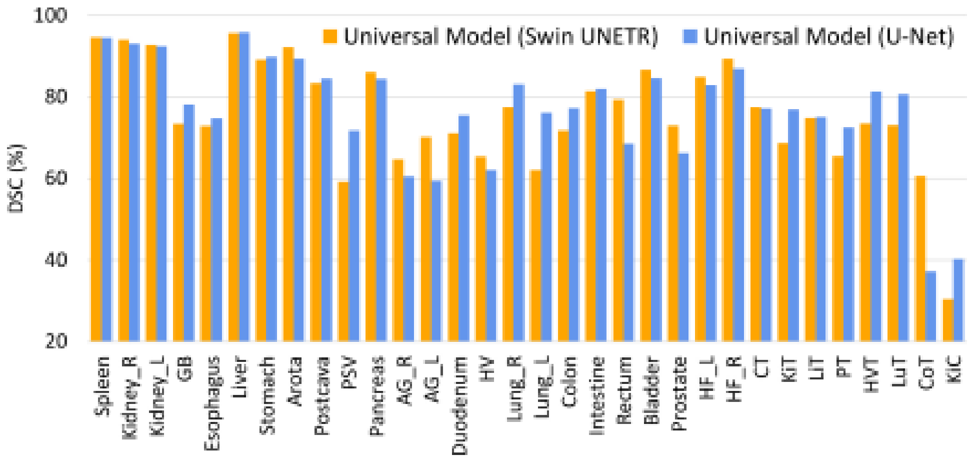


Figure 8. ***Expansibility:* flexible backbones.** Universal Model can be expanded to CNN-based (*e.g.*, U-Net [55]) and Transformer-based (*e.g.*, Swin UNETR [64]) backbone. For the abbreviation of some organs, please refer to Appendix Table 14. Both backbones achieve comparable results.

Experiments

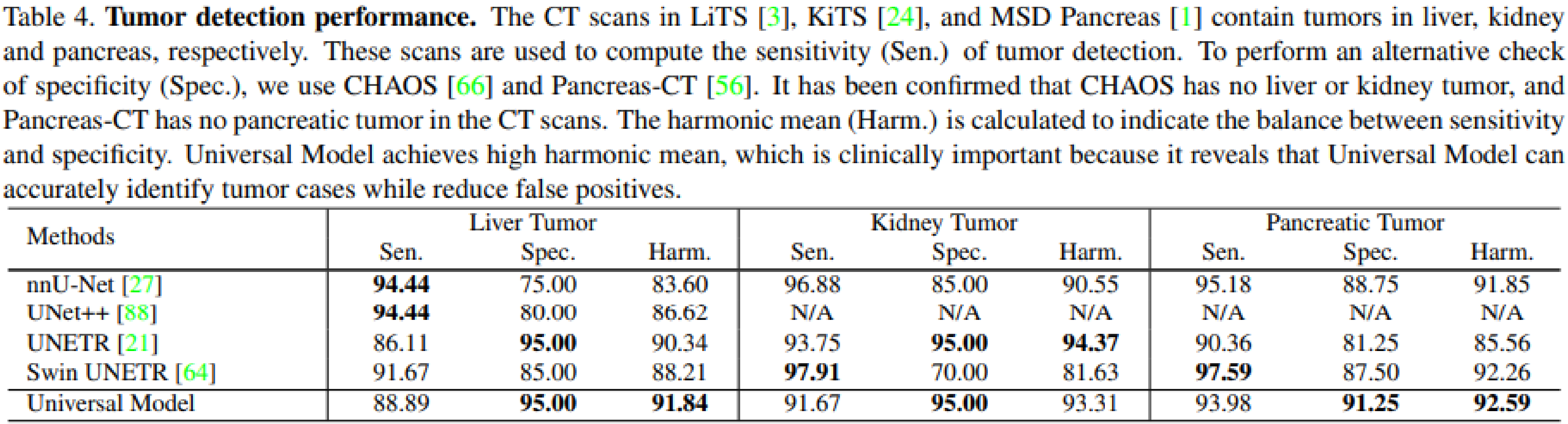


Table 4. **Tumor detection performance.** The CT scans in LiTS [3], KiTS [24], and MSD Pancreas [1] contain tumors in liver, kidney and pancreas, respectively. These scans are used to compute the sensitivity (Sen.) of tumor detection. To perform an alternative check of specificity (Spec.), we use CHAOS [66] and Pancreas-CT [56]. 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 [27]	**94.44**	75.00	83.60	96.88	85.00	90.55	95.18	88.75	91.85
UNet++ [88]	**94.44**	80.00	86.62	N/A	N/A	N/A	N/A	N/A	N/A
UNETR [21]	86.11	**95.00**	90.34	93.75	**95.00**	**94.37**	90.36	81.25	85.56
Swin UNETR [64]	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**

Experiments

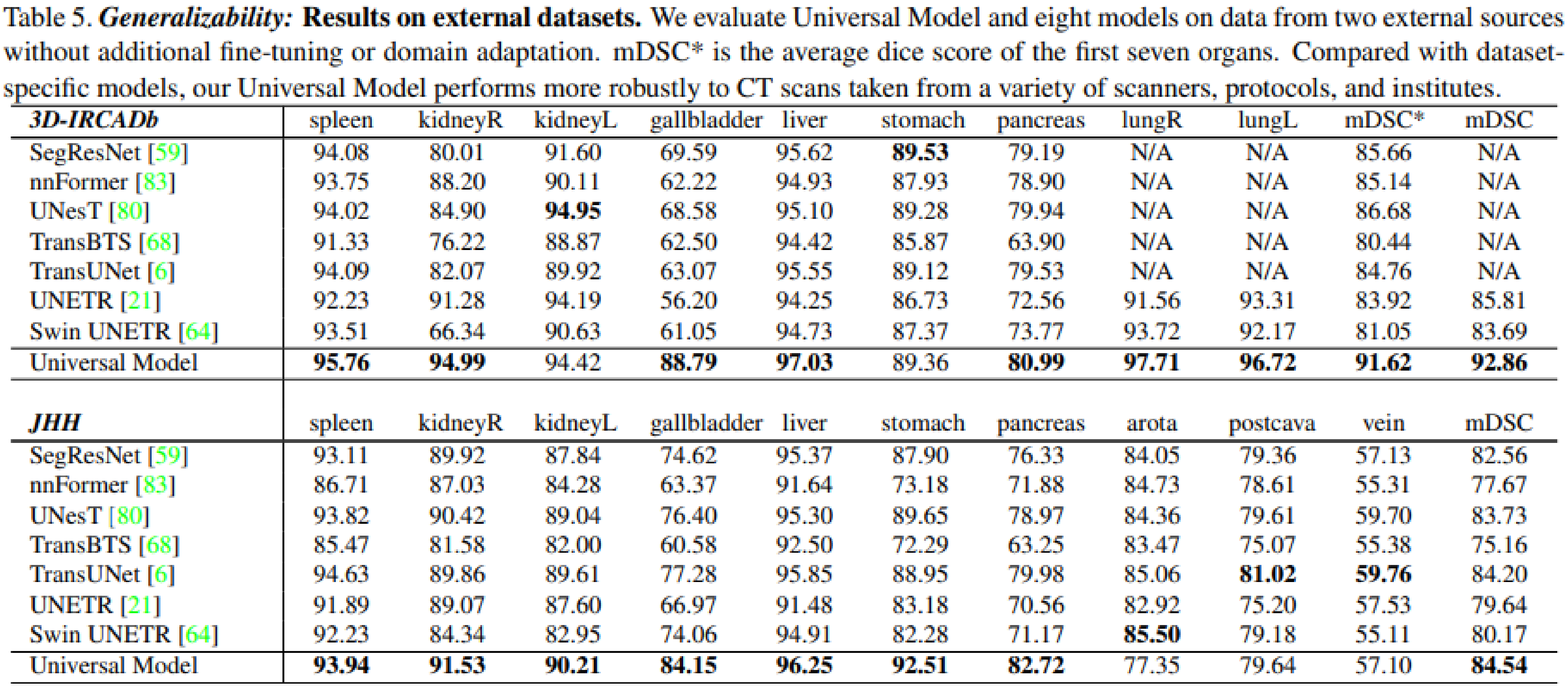


Table 5. ***Generalizability:* Results on external datasets.** We evaluate Universal Model and eight models on data from two external sources without additional fine-tuning or domain adaptation. mDSC* is the average dice score of the first seven organs. Compared with dataset-specific models, our Universal 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 [59]	94.08	80.01	91.60	69.59	95.62	**89.53**	79.19	N/A	N/A	85.66	N/A
nnFormer [83]	93.75	88.20	90.11	62.22	94.93	87.93	78.90	N/A	N/A	85.14	N/A
UNesT [80]	94.02	84.90	**94.95**	68.58	95.10	89.28	79.94	N/A	N/A	86.68	N/A
TransBTS [68]	91.33	76.22	88.87	62.50	94.42	85.87	63.90	N/A	N/A	80.44	N/A
TransUNet [6]	94.09	82.07	89.92	63.07	95.55	89.12	79.53	N/A	N/A	84.76	N/A
UNETR [21]	92.23	91.28	94.19	56.20	94.25	86.73	72.56	91.56	93.31	83.92	85.81
Swin UNETR [64]	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 [59]	93.11	89.92	87.84	74.62	95.37	87.90	76.33	84.05	79.36	57.13	82.56
nnFormer [83]	86.71	87.03	84.28	63.37	91.64	73.18	71.88	84.73	78.61	55.31	77.67
UNesT [80]	93.82	90.42	89.04	76.40	95.30	89.65	78.97	84.36	79.61	59.70	83.73
TransBTS [68]	85.47	81.58	82.00	60.58	92.50	72.29	63.25	83.47	75.07	55.38	75.16
TransUNet [6]	94.63	89.86	89.61	77.28	95.85	88.95	79.98	85.06	**81.02**	**59.76**	84.20
UNETR [21]	91.89	89.07	87.60	66.97	91.48	83.18	70.56	82.92	75.20	57.53	79.64
Swin UNETR [64]	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**

Experiments

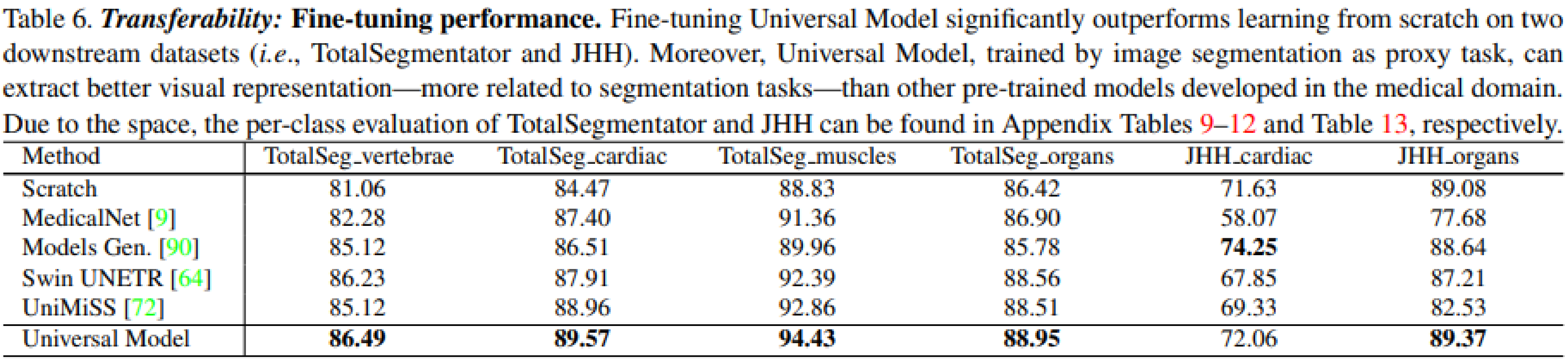


Table 6. ***Transferability:* Fine-tuning performance.** Fine-tuning Universal Model significantly outperforms learning from scratch on two downstream datasets (*i.e.*, TotalSegmentator and JHH). Moreover, Universal Model, trained by image segmentation as proxy task, can extract better visual representation—more 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 [9]	82.28	87.40	91.36	86.90	58.07	77.68
Models Gen. [90]	85.12	86.51	89.96	85.78	**74.25**	88.64
Swin UNETR [64]	86.23	87.91	92.39	88.56	67.85	87.21
UniMiSS [72]	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**

Experiments – Ablations

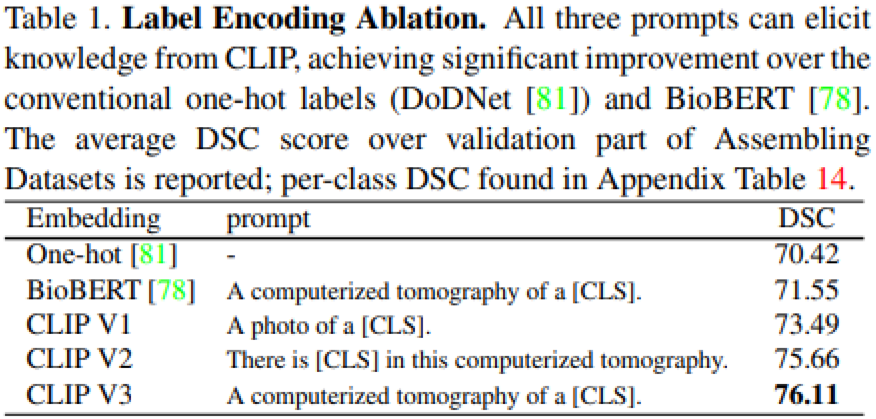
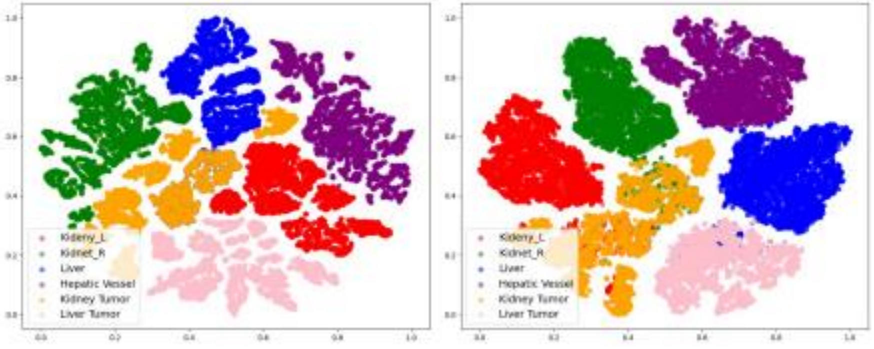


Table 1. **Label Encoding Ablation.** All three prompts can elicit knowledge from CLIP, achieving significant improvement over the conventional one-hot labels (DoDNet [81]) and BioBERT [78]. The average DSC score over validation part of Assembling Datasets is reported; per-class DSC found in Appendix Table 14.

Embedding	prompt	DSC
One-hot [81]	-	70.42
BioBERT [78]	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**



(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 selected 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.

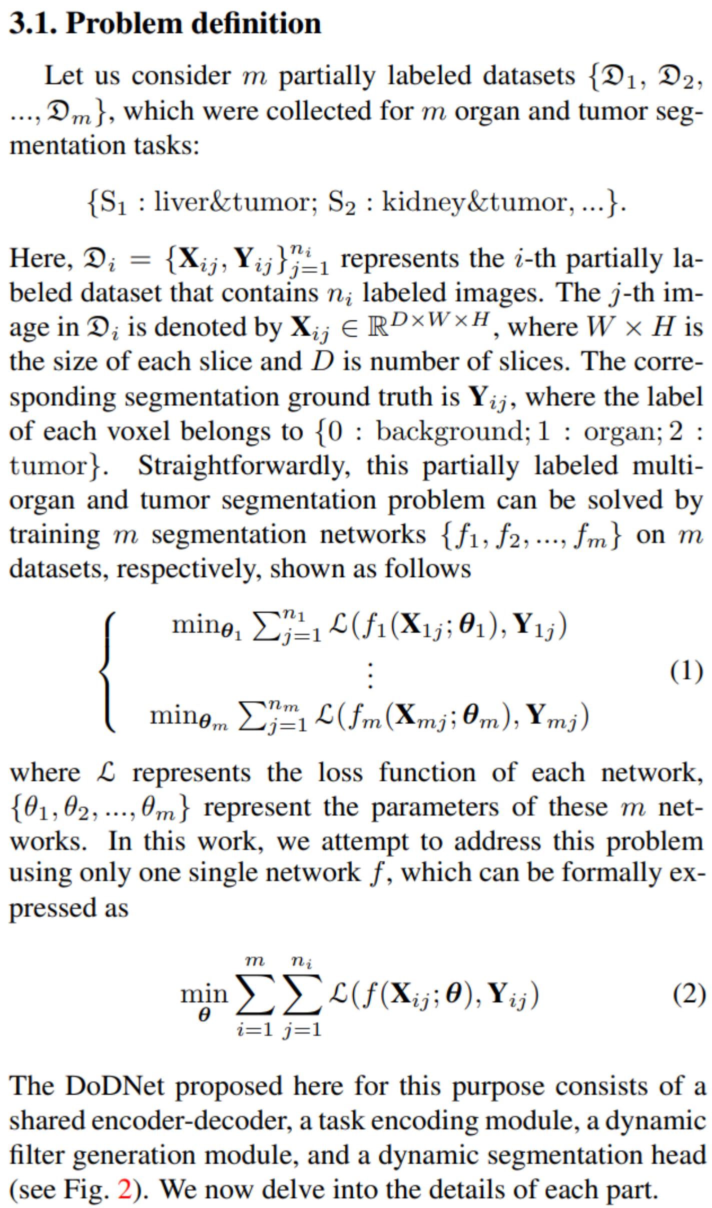
- 이 연구에서는 복부 기관 분할과 종양 탐지를 위한 CLIP 기반 범용 모델을 제시합니다.
- 라벨의 불일치와 직교성 문제를 해결하기 위해, CLIP 임베딩을 분할 모델과 통합하여 유연하면서 강력한 분할 도구를 만들었습니다.
- 이 모델은 부분적으로 라벨링된 데이터셋에서 효과적으로 학습하고, MSD와 BTCV에서 모두 첫 번째로 랭크되는 등 높은 성능을 달성하였습니다.
- 여섯 개의 기관의 분할 정확도는 사람의 수준에 도달하였습니다.
- 중요하게도, 우리의 연구는 CLIP 임베딩이 일반적으로 사용되는 원-핫 임베딩보다 기관과 종양 간의 의미 있는 해부학적 관계를 더 강력하게 구축할 수 있음을 보여줍니다.
- 또한, 실험 결과를 통해 CLIP 기반 범용 모델의 여러 중요한 임상적 이점, 즉 뛰어난 효율성, 일반화 가능성, 이식성, 확장성을 확인하였습니다.

Reference

- https://openaccess.thecvf.com/content/CVPR2021/papers/Zhang_DoDNet_Learning_To_Segment_Multi-Organ_and_Tumors_From_Multiple_Partially_CVPR_2021_paper.pdf

- https://arxiv.org/pdf/2301.00785.pdf

Appendix #1



3.1. Problem definition

Let us consider m partially labeled datasets $\{\mathfrak{D}_1, \mathfrak{D}_2, ..., \mathfrak{D}_m\}$, which were collected for m organ and tumor segmentation tasks:

$$\{\mathrm{S}_1 : \text{liver\&tumor}; \; \mathrm{S}_2 : \text{kidney\&tumor}, ...\}.$$

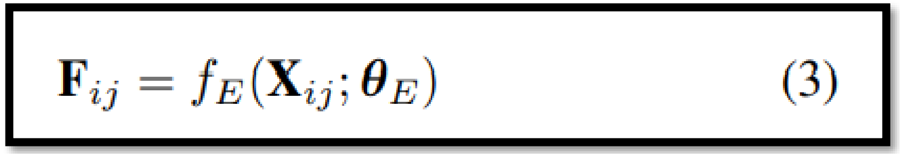
Here, $\mathfrak{D}_i = \{\mathbf{X}_{ij}, \mathbf{Y}_{ij}\}_{j=1}^{n_i}$ represents the i-th partially labeled dataset that contains n_i labeled images. The j-th image in $\mathfrak{D}_i$ is denoted by $\mathbf{X}_{ij} \in \mathbb{R}^{D \times W \times H}$, where $W \times H$ is the size of each slice and D is number of slices. The corresponding segmentation ground truth is $\mathbf{Y}_{ij}$, where the label of each voxel belongs to $\{0 : \mathrm{background}; 1 : \mathrm{organ}; 2 : \mathrm{tumor}\}$. Straightforwardly, this partially labeled multi-organ and tumor segmentation problem can be solved by training m segmentation networks $\{f_1, f_2, ..., f_m\}$ on m datasets, respectively, shown as follows

$$\begin{cases} \min_{\boldsymbol{\theta}_1} \sum_{j=1}^{n_1} \mathcal{L}(f_1(\mathbf{X}_{1j}; \boldsymbol{\theta}_1), \mathbf{Y}_{1j}) \\ \vdots \\ \min_{\boldsymbol{\theta}_m} \sum_{j=1}^{n_m} \mathcal{L}(f_m(\mathbf{X}_{mj}; \boldsymbol{\theta}_m), \mathbf{Y}_{mj}) \end{cases} \tag{1}$$

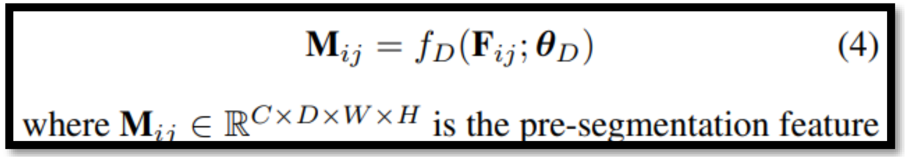
where $\mathcal{L}$ represents the loss function of each network, $\{\theta_1, \theta_2, ..., \theta_m\}$ represent the parameters of these m networks. In this work, we attempt to address this problem using only one single network f, which can be formally expressed as

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathcal{L}(f(\mathbf{X}_{ij}; \boldsymbol{\theta}), \mathbf{Y}_{ij}) \tag{2}$$

The DoDNet proposed here for this purpose consists of a shared encoder-decoder, a task encoding module, a dynamic filter generation module, and a dynamic segmentation head (see Fig. 2). We now delve into the details of each part.

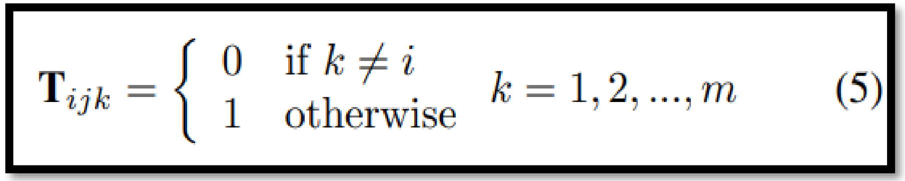


$$\mathbf{F}_{ij} = f_E(\mathbf{X}_{ij}; \boldsymbol{\theta}_E) \tag{3}$$



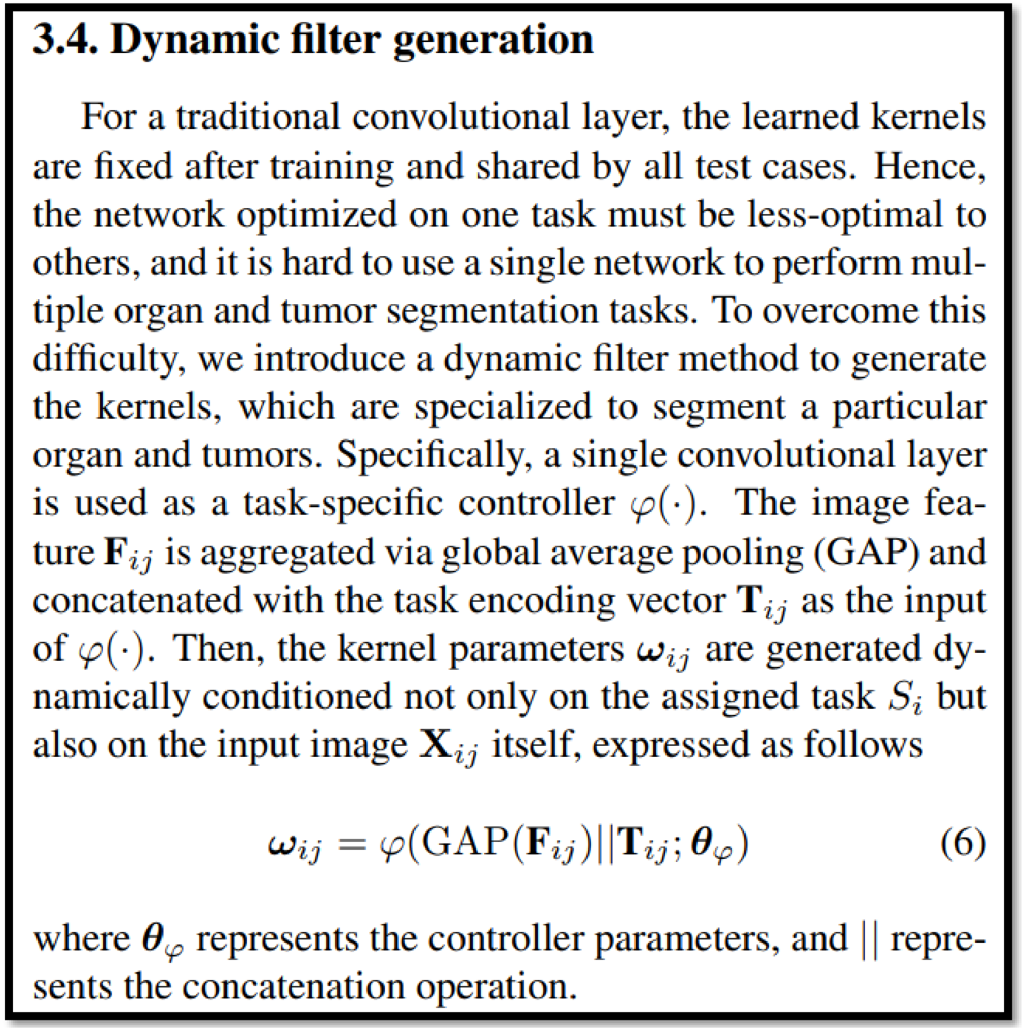
$$\mathbf{M}_{ij} = f_D(\mathbf{F}_{ij}; \boldsymbol{\theta}_D) \tag{4}$$

where $\mathbf{M}_{ij} \in \mathbb{R}^{C \times D \times W \times H}$ is the pre-segmentation feature



$$\mathbf{T}_{ijk} = \begin{cases} 0 & \text{if } k \neq i \\ 1 & \text{otherwise} \end{cases} \quad k = 1, 2, ..., m \tag{5}$$

Appendix #2

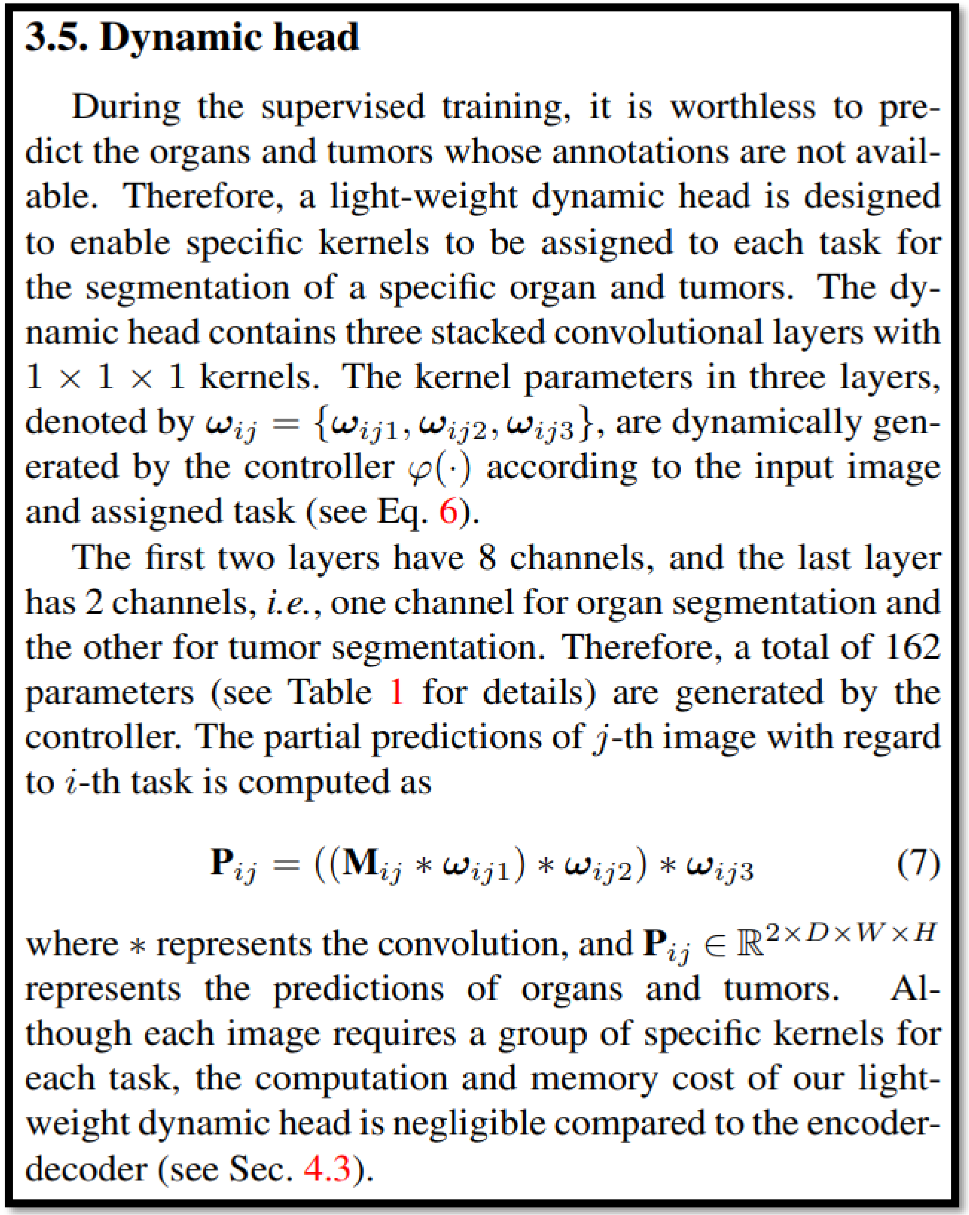


3.4. Dynamic filter generation

For a traditional convolutional layer, the learned kernels are fixed after training and shared by all test cases. Hence, the network optimized on one task must be less-optimal to others, and it is hard to use a single network to perform multiple organ and tumor segmentation tasks. To overcome this difficulty, we introduce a dynamic filter method to generate the kernels, which are specialized to segment a particular organ and tumors. Specifically, a single convolutional layer is used as a task-specific controller $\varphi(\cdot)$. The image feature $\mathbf{F}_{ij}$ is aggregated via global average pooling (GAP) and concatenated with the task encoding vector $\mathbf{T}_{ij}$ as the input of $\varphi(\cdot)$. Then, the kernel parameters $\boldsymbol{\omega}_{ij}$ are generated dynamically conditioned not only on the assigned task S_i but also on the input image $\mathbf{X}_{ij}$ itself, expressed as follows

$$\boldsymbol{\omega}_{ij} = \varphi(\mathrm{GAP}(\mathbf{F}_{ij})||\mathbf{T}_{ij}; \boldsymbol{\theta}_{\varphi}) \tag{6}$$

where $\boldsymbol{\theta}_{\varphi}$ represents the controller parameters, and $||$ represents the concatenation operation.



3.5. Dynamic head

During the supervised training, it is worthless to predict the organs and tumors whose annotations are not available. Therefore, a light-weight dynamic head is designed to enable specific kernels to be assigned to each task for the segmentation of a specific organ and tumors. The dynamic head contains three stacked convolutional layers with $1 \times 1 \times 1$ kernels. The kernel parameters in three layers, denoted by $\boldsymbol{\omega}_{ij} = \{\boldsymbol{\omega}_{ij1}, \boldsymbol{\omega}_{ij2}, \boldsymbol{\omega}_{ij3}\}$, are dynamically generated by the controller $\varphi(\cdot)$ according to the input image and assigned task (see Eq. 6).

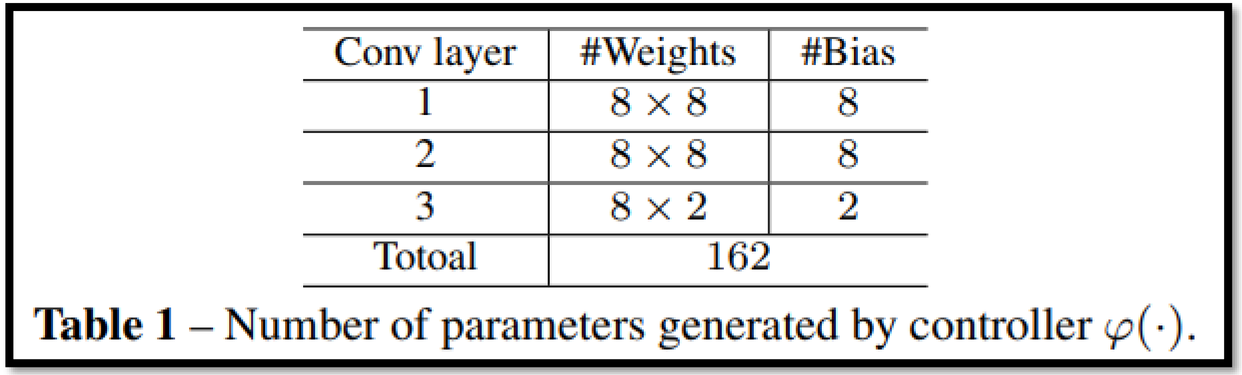
The first two layers have 8 channels, and the last layer has 2 channels, *i.e.*, one channel for organ segmentation and the other for tumor segmentation. Therefore, a total of 162 parameters (see Table 1 for details) are generated by the controller. The partial predictions of j-th image with regard to i-th task is computed as

$$\mathbf{P}_{ij} = ((\mathbf{M}_{ij} * \boldsymbol{\omega}_{ij1}) * \boldsymbol{\omega}_{ij2}) * \boldsymbol{\omega}_{ij3} \tag{7}$$

where $*$ represents the convolution, and $\mathbf{P}_{ij} \in \mathbb{R}^{2 \times D \times W \times H}$ represents the predictions of organs and tumors. Although each image requires a group of specific kernels for each task, the computation and memory cost of our light-weight dynamic head is negligible compared to the encoder-decoder (see Sec. 4.3).

# PaLM-E: An Embodied Multimodal Language Model

Appendix #3



Conv layer	#Weights	#Bias
1	8×8	8
2	8×8	8
3	8×2	2
Totoal	162	

Table 1 – Number of parameters generated by controller $\varphi(\cdot)$.

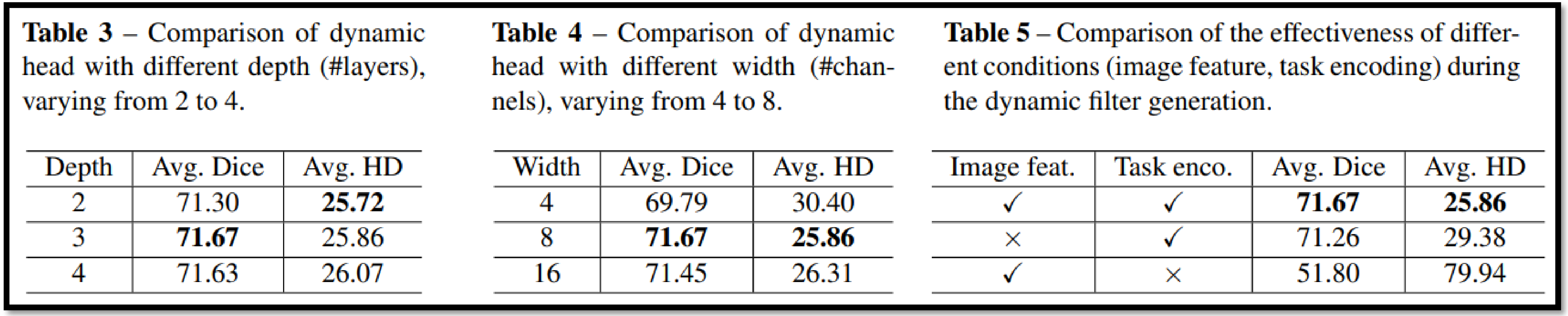


Table 3 – Comparison of dynamic head with different depth (#layers), varying from 2 to 4.

Depth	Avg. Dice	Avg. HD
2	71.30	**25.72**
3	**71.67**	25.86
4	71.63	26.07

Table 4 – Comparison of dynamic head with different width (#channels), varying from 4 to 8.

Width	Avg. Dice	Avg. HD
4	69.79	30.40
8	**71.67**	**25.86**
16	71.45	26.31

Table 5 – Comparison of the effectiveness of different conditions (image feature, task encoding) during the dynamic filter generation.

Image feat.	Task enco.	Avg. Dice	Avg. HD
✓	✓	**71.67**	**25.86**
×	✓	71.26	29.38
✓	×	51.80	79.94

# PaLM-E: An Embodied Multimodal Language Model

Thank you for your Attention….!