Appendix: CLIP-Driven Universal Model

Abstract. In this supplementary material, we provide additional information about the CLIP-Driven Universal Model and the assembly of 14 public datasets, as well as more detailed experimental results than those in the main paper. Appendix A discusses the influence of the medical prompt template. Appendix B provides the specifications for the assembly of datasets. Appendix C elaborates on the implementation details, including the data augmentations, model network structures and evaluation metrics used in the main paper. Appendix D supplements the qualitative and quantitative analysis in the main paper, including the visualization of kidney tumors and liver tumors, complete evaluation results of the transfer learning experiment, and whole embedding space visualization. Finally, Appendix E visualizes several open challenges when assembling public datasets with partial labels.

A. Medical Prompt Template

To fully explore the effect of templates on CLIP embedding, an experiment is performed in the whole assembly of datasets as shown in Table 1. Four text templates are employed to show the context, *i.e.*, "V1: A computerized tomography of a [CLS].", "V2: There is [CLS] in this computerized tomography.", "V3: This computerized tomography has a [CLS].", "V4: A photo of a [CLS].". The effectiveness of the prompt template is slightly different from the toy experiment. With increasing organ numbers, templates V1 and V2 still show better performance in encoding the relationship, but template V3 would deteriorate the results. In addition, a widely used template V4 could also promote the segmentation performance.

As known, the prompt template is a crucial factor for text model [92, 44]. How select an appropriate template is still an open problem for the medical image text-vision models. We encourage more future work to explore this area.

B. Assembly of Datasets

The assembly of datasets consists of 14 publicly available datasets for training and 2 public datasets and 1 large-scale private dataset for testing (summarized in Table 7). It is non-trivial to assemble datasets annotated from various institutions since the annotation protocols are inconsistent. As mentioned in the main paper, we unify the label index for all datasets. The corresponding relationship is as follows. (Spleen, 1); (Right Kidney, 2); (Left Kidney, 3); (Gall Bladder, 4); (Esophagus, 5); (Liver, 6); (Stomach, 7); (Aorta, 8); (Postcava, 9); (Portal Vein and Splenic Vein, 10); (Pancreas, 11); (Right Adrenal Gland, 12); (Left Adrenal Gland, 13); (Duodenum, 14); (Hepatic Vessel, 15); (Right Lung, 16); (Left Lung, 17); (Colon, 18); (Intestine, 19); (Rectum, 20); (Bladder, 21); (Prostate/Uterus, 22); (Head of Femur Left, 23); (Head of Femur Right, 24); (Celiac Truck, 25); (Kidney Tumor, 26); (Liver Tumor, 27); (Pancreas Tumor, 28); (Hepatic Vessel Tumor, 29); (Lung Tumor, 30); (Colon Tumor, 31); (Kidney Cyst, 32). Firstly, we map all the datasets into the standard index template. Then, for these datasets (KiTS, WORD, AbdomenCT-1K, and CT-ORG), which do not distinguish between the left and right organs, we split the organ (Kidney, Adrenal Gland, and Lung) into left part and right part through the script. In addition, we have taken the inclusion relation into consideration, *e.g.*, the organ tumor is part of the organ, and the hepatic vessel is inside the liver. Since we formulate each organ segmentation result as a binary mask, we can organize the segmentation ground truth for these overlapped organs independently in a binary mask manner.

Pancreas-CT [60] consists of 82 contrast-enhanced abdominal CT volumes. This dataset only provides the pancreas label annotated by an experienced radiologist, and all CT scans have no pancreatic tumor.

LiTS [3] contains 131 and 70 contrast-enhanced 3-D abdominal CT scans for training and testing, respectively. The data set was acquired by different scanners and protocols at six different clinical sites, with a largely varying in-plane resolution from 0.55 to 1.0 mm and slice spacing from 0.45 to 6.0 mm.

KiTS [24] includes 210 training cases and 90 testing cases with annotations provided by the University of Minnesota Medical Center. Each CT scan has one or more kidney tumors.

AbdomenCT-1K [43] consists of 1112 CT scans from five datasets with liver, kidney, spleen, and pancreas annotations.

CT-ORG [58] is composed of 140 CT images containing 6 organ classes, which are from 8 different medical centers. Most of the images exhibit liver lesions, both benign and malignant.

CHAOS [71] provides 20 patients for multi-organ segmentation. All CT scans have no liver tumor.

MSD CT Tasks [1] includes liver, lung, pancreas, colon, hepatic vessel, and spleen tasks for a total of 947 CT scans with 4 organs and 5 tumors.

BTCV [36] consists of 50 abdominal CT scans from metastatic liver cancer patients or post-operative ventral hernia patients. They are collected from the Vanderbilt University Medical Center.

AMOS22 [31] is the abbreviation of the multi-modality abdominal multi-organ segmentation challenge of 2022. The AMOS dataset contains 500 CT with voxel-level annotations of 15 abdominal organs.

WORD [42] collects 150 CT scans from 150 patients before the radiation therapy in a single center. All of them are

Table 7. **The information for an assembly of datasets.** We have developed a *Universal Model* from an assembly of 1–14 public datasets. The official test and validation sets of Medical Segmentation Decathlon (MSD) and Beyond the Cranial Vault (BTCV) are used to benchmark the performance of organ segmentation (§4.1) and tumor detection (§4.2). 3D-IRCADb (15), TotalSegmentator (16) and a large-scale private dataset (17), consisting of 5,038 CT scans with 21 annotated organs, are used for independent evaluation of model generalizability and transferability (§5). This list will continue to grow when more annotated datasets become available.

Datasets	# Targets	# Scans	Annotated Organs or Tumors
1. Pancreas-CT [60]	1	82	Pancreas
2. LiTS [3]	2	201	Liver, Liver Tumor*
3. KiTS [24]	2	300	Kidney, Kidney Tumor*
4. AbdomenCT-1K [43]	4	1,000	Spleen, Kidney, Liver, Pancreas
5. CT-ORG [58]	4	140	Lung, Liver, Kidneys and Bladder
6. CHAOS [71]	4	40	Liver, Left Kidney, Right Kidney, Spl
7-11. MSD CT Tasks [1]	9	947	Spl, Liver and Tumor*, Lung Tumor*, Colon Tumor*, Pan and Tumor*, Hepatic Vessel and Tumor*
12. BTCV [36]	13	50	Spl, RKid, LKid, Gall, Eso, Liv, Sto, Aor, IVC, R&SVeins, Pan, RAG, LAG
13. AMOS22 [31]	15	500	Spl, RKid, LKid, Gall, Eso, Liv, Sto, Aor, IVC, Pan, RAG, LAG, Duo, Bla, Pro/UTE
14. WORD [42]	16	150	Spl, RKid, LKid, Gall, Eso, Liv, Sto, Pan, RAG, Duo, Col, Int, Rec, Bla, LFH, RFH
15. 3D-IRCADb [65]	13	20	Liv, Liv Cyst, RLung, LLung, Venous, PVein, Aor, Spl, RKid, LKid, Gall, IVC
16. TotalSegmentator [76]	104	1,024	Clavicula, Humerus, Scapula, Rib 1-12, Vertebrae C1-7, Vertebrae T1-9, Vertebrae L1-5, Hip, Sacrum, Femur, Aorta, Pulmonary Artery, Right Ventricle, Right Atrium, Left Atrium, Left Ventricle, Myocardium, PVein, SVein, IVC, Iliac Artery, Iliac Vena, Brain, Trachea, Lung Upper Lobe, Lung Middle Lobe, Lung Lower Lobe, AG, Spl, Liv, Gall, Pan, Kid, Eso, Sto, Duo, Small Bowel, Colon, Bla, Autochthon, Iliopsoas, Gluteus Minimus, Gluteus Medius, Gluteus Maximus
17. JHH (*private*)	21	5,038	Aor, AG, CBD, Celiac AA, Colon, duo, Gall, IVC, Lkid, RKid, Liv, Pan, Pan Duct, SMA, Small bowel, Spl, Sto, Veins, Kid LtRV, Kid RtRV, CBD Stent, PDAC*, PanNET*, Pancreatic Cyst*

scanned by a SIEMENS CT scanner without appearance enhancement. Each CT volume consists of 159 to 330 slices of 512 × 512 pixels.

3D-IRCADb [65] contains 20 venous phase enhanced CT scans. Each CT scan has various annotations, and only annotated organs are tested to validate the model's generalizability.

TotalSegmentator [76] collects 1024 CT scans randomly sampled from PACS over the timespan of the last 10 years. The dataset contains CT images with different sequences (native, arterial, portal venous, late phase, dual-energy), with and without contrast agent, with different bulb voltages, with different slice thicknesses and resolution and with different kernels (soft tissue kernel, bone kernel).

JHH (*private*) contains 5038 CT scans with 21 annotated organs, where each case was scanned by contrast-enhanced CT in both venous and arterial phases, acquired on Siemens MDCT scanners. The JHH dataset is used to investigate the extensibility of new classes.

C. Implementation Details

C.1. Data Augmentation

Our data augmentation is implemented in python with MONAI[10]. The orientation of CT scans is changed into specified axcodes. Isotropic spacing is adopted to re-slice each scan to the same voxel size of $1.5 \times 1.5 \times 1.5 mm^3$. We truncate the intensity in each scan to the range $[-175, 250]$ and linearly normalize them to $[0, 1]$. Considering the valid part is part of the whole medical image, we crop only the foreground object based on the images. During training, we crop random fixed-sized $96 \times 96 \times 96$ regions with the center being a foreground or background voxel based on the predefined ratio. Also, we randomly rotate the input patch by 90 degrees and shift intensity with 0.1 offset with 0.1 and 0.2 probability. To avoid confusion between the organ in the right and left parts, we do not use mirroring augmentation.

Table 8. **The 5-fold cross-validation performance on MSD.** These are the tabular comparison between Universal Model and Swin UNETR [68] (previously ranked first on the MSD leaderboard). The performance is evaluated by DSC scores.

Task		SwinUNETR [68]	Ours
Task 03	Liver	94.12±2.34	**96.49±0.23**
	Liver Tumor	57.86±4.72	**71.94±3.74**
Task 06	Lung Tmuor	**68.90±5.44**	67.15±5.81
Task 07	Pancreas	80.06±0.83	**82.70±1.96**
	Panc. Tumor	52.53±3.76	**60.82±10.2**
Task 08	Hepat. Ves.	62.33±2.44	**62.55±3.64**
	Ves. Tumor	68.56±3.82	**69.39±2.29**
Task 09	Spleen	95.80±0.56	**96.71±0.21**
Task 10	Col. Tumor	50.45±10.1	**62.14±17.8**

C.2. Network Structures

Text branch. We apply the pre-trained text encoder "ViT-B/32" of the CLIP as the text branch[11]. We can extract and store the text features to reduce overhead brought by the text encoder in the training and inference stage since the CLIP embedding only depends on the dictionary, which is fixed.

[10] https://monai.io/

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

Vision branch. We adopt Swin UNETR as a vision encoder. The Swin UNETR consists of 4 attention stages comprising 2 transformer blocks and 5 convolution stages comprising of CNN-based structure. In the attention stage, a patch merging layer is used to reduce the resolution by a factor of 2. Stage 1 consists of a linear embedding layer and transformer blocks that maintain the number of tokens as $\frac{H}{2} \times \frac{W}{2} \times \frac{D}{2}$. a patch merging layer groups patches with resolution $2 \times 2 \times 2$ and concatenates them, resulting in a 4C-dimensional feature embedding. A linear layer is then used to down-sample the resolution by reducing the dimension to 2C. The same procedure continues in stages 2, 3, and 4 [68]. The text-based controller is a single convolutional layer, which takes the CLIP embedding and global pooling feature from the last convolution stages in the vision encoder as input.

C.3. Evaluation Metrics

The Dice similarity coefficient (DSC) and Normalized Surface Distance (NSD) are used as measurements for 3D segmentation results. The DSC metric is defined as:

$$\text{DSC} = \frac{2\sum_{i=1}^{I} Y_i \hat{Y}_i}{\sum_{i=1}^{I} Y_i + \sum_{i=1}^{I} \hat{Y}_i}, \tag{1}$$

where Y and $\hat{Y}$ denote the ground truth and prediction of voxel values. The details of Normalized Surface Distance (NSD) could refer to Sec. 4.6 in [49].

D. Additional Evaluations

Table 8 shows the detailed numerical result between Universal Model and Swin UNETR. Tables 9–12 and Table 13 show the per-class evaluation of TotalSegmentator and JHH, which validates the transferability of the proposed Universal Model.

Figure 9 exhibits the contour line comparison among Universal Model and two human experts. We can see the model predictions are roughly similar to human annotation, which validates the effectiveness of the pseudo label generated by our Universal Model.

Figure 11 and Figure 10 shows several kidney and liver tumor cases comparison among the proposed Universal Model and four competitive baseline methods. Our method can not only detect small and big tumors in various organs but also not generate false positives of tumors.

Table 14 shows the ablation study results of CLIP embedding, which is an extension for Table 1. Dice scores for each organ and tumor are reported.

Figure 12 shows the whole embedding space of baseline method and universal model. Our method shows better semantic relationship of anatomical structure.

E. Discussion of Open Challenges

Inconsistent label protocols. The first open challenge is the inconsistent annotation protocol. The annotation standard is different from institution to institution. In AMOS, "Aorta" refers to the entire region of Aorta, but in AbdomenCT-1K, a part of the upper regions annotation is missing. It is because of the inconsistent definitions in different datasets and this requires considerable manual corrections of several experienced radiology experts when assembling these datasets together.

Long-tail problem. The assembly of public datasets leads to severe class imbalance problems, especially for small tumors. We count the proportion of each organ and tumor in Figure 15. The assembly of datasets has a severe long-tail distribution, which would lead to unsatisfactory performance of tumor classes. Mitigating the long-tail distribution would contribute to more robust detection of the tumor. In this paper, we utilize data augmentation to alleviate the long-tail problem, but more research is encouraged to explore the solution to these two problems.

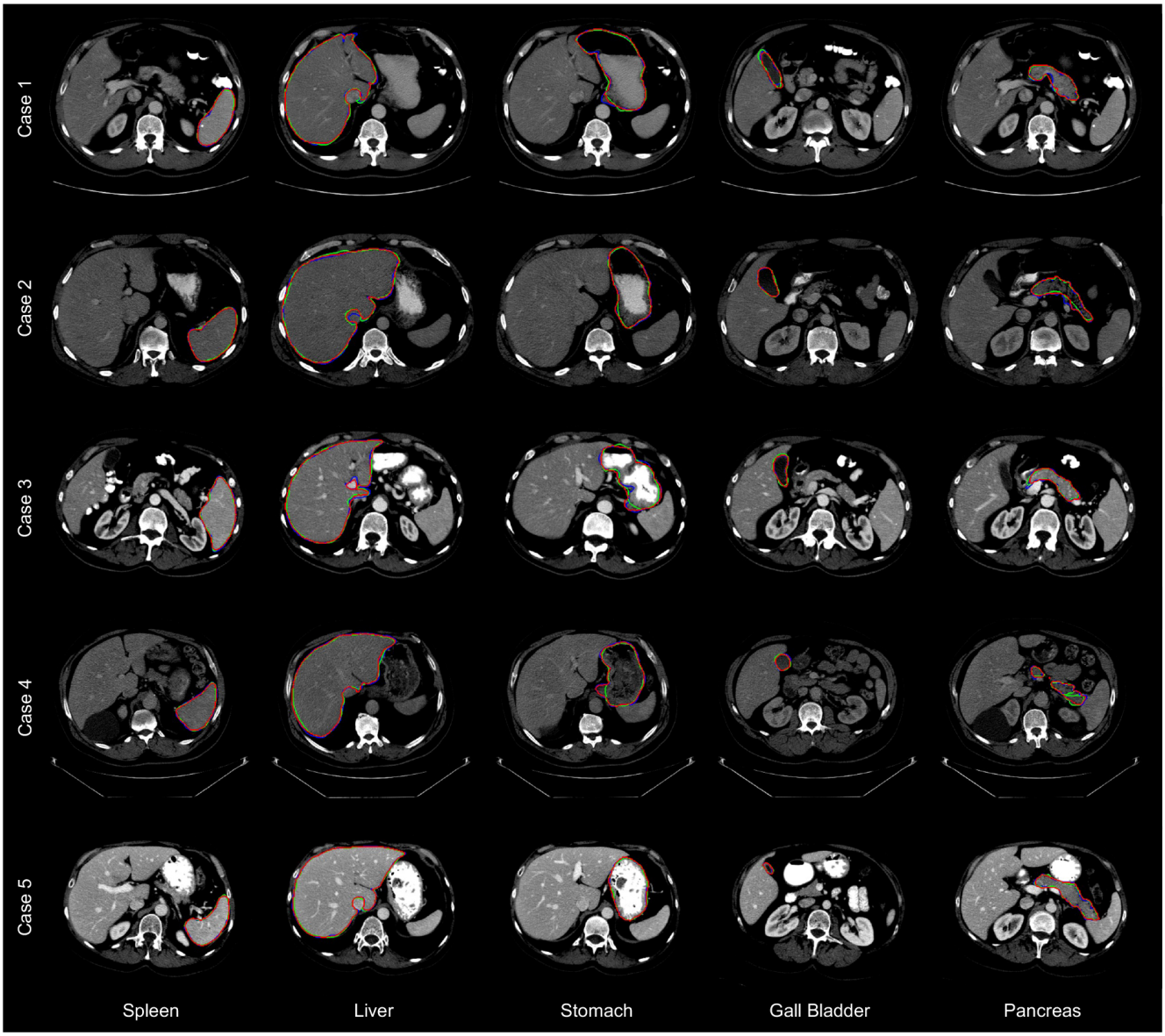

Figure 9. **Contour line comparison among pseudo labels and two human experts.** The red line represents the annotation from Doctor 1; green line indicates the annotation from Doctor 2; blue line shows the results generated by Universal Model. Examples of CT scans annotated by our pseudo labels and two human experts with contour line comparison. The prediction results of these organs generated by the medical model are comparable with human experts.

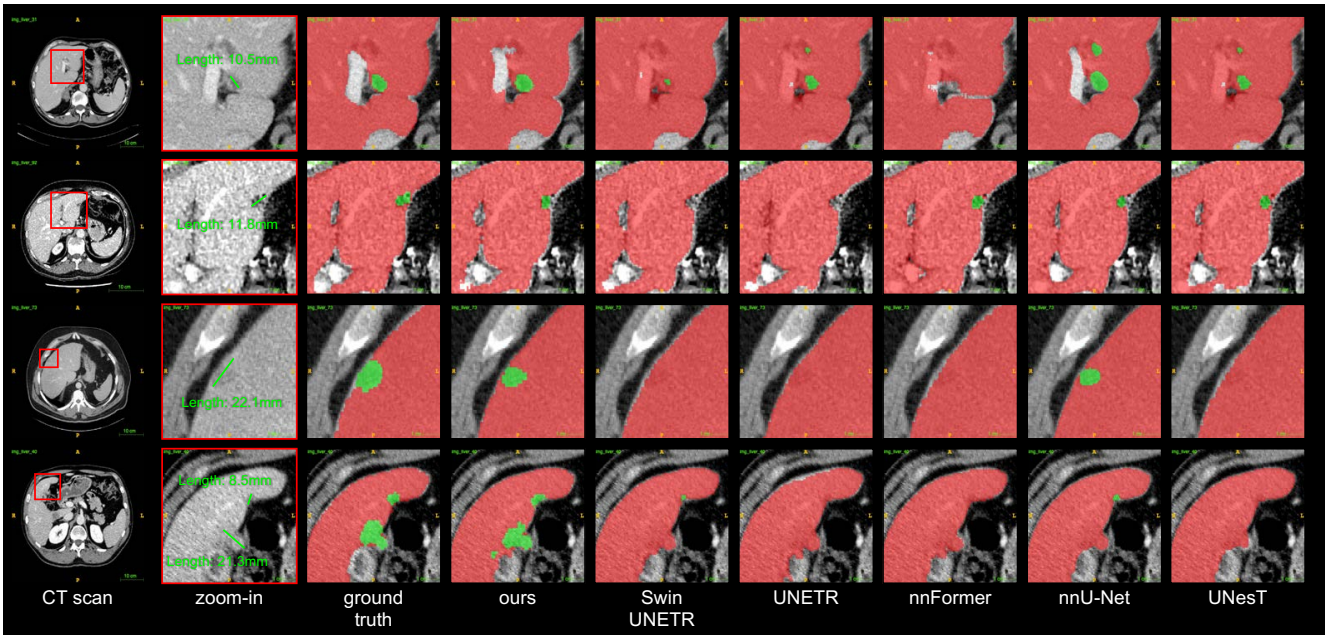

Figure 10. **Liver tumor detection.** Qualitative visualizations of the proposed Universal Model and four competitive baseline methods. We review the detection results of tumors from smaller to larger sizes (Rows 1–4). The Universal Model succeeds in detecting small tumors ignored by other methods and in detecting multiple tumors in one CT. In addition, it avoids the false positive prediction, which validates the good practicability of Universal Model.

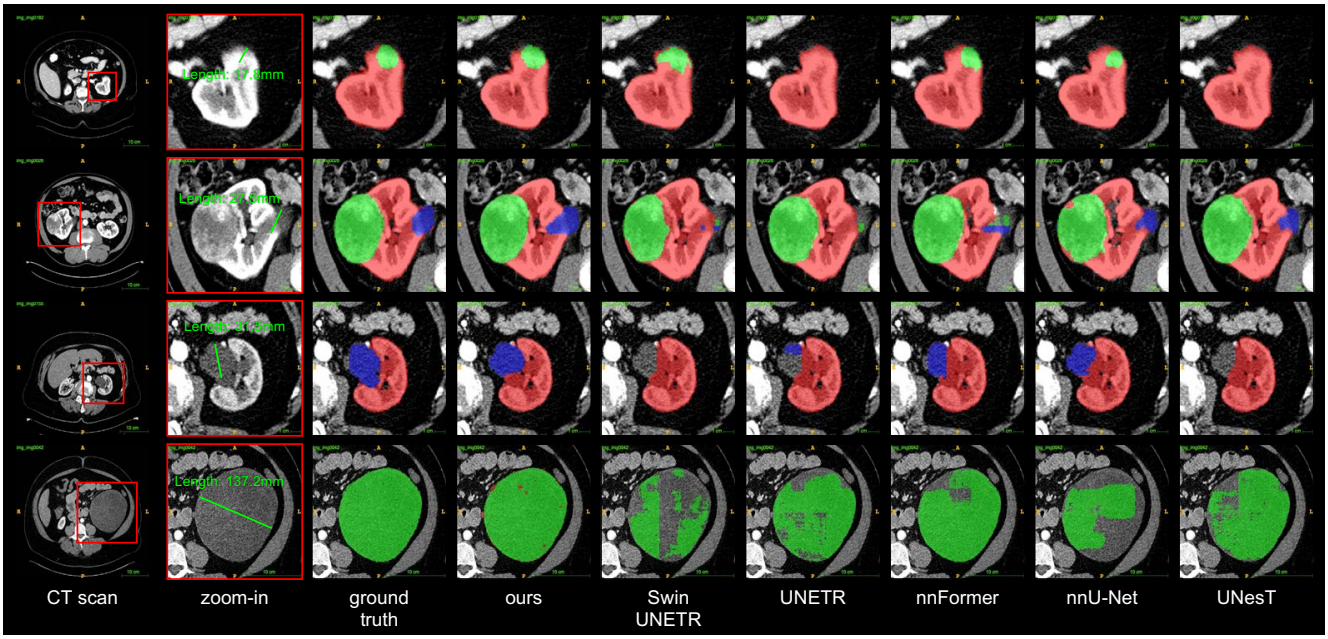

Figure 11. **Kidney tumor detection.** Qualitative visualizations of the proposed Universal Model and four competitive baseline methods. We review the detection results of tumors from smaller to larger sizes (Rows 1–4). The Universal Model can detect well not only on the kidneys (red region), but also kidney tumors (green region) and cysts (blue region).

Table 9. **The complete evaluation of TotalSeg_vertebrae.** The results are evaluated by DSC. Our Universal Model represents the best transferability.

Method	L5	L4	L3	L2	L1	T12	T11	T10	T9	T8	T7	T6
Scratch	86.68	88.37	89.83	84.28	91.98	87.45	88.29	86.78	83.50	75.70	77.73	75.84
MedicalNet [10]	**91.72**	91.01	86.03	84.73	91.52	**89.98**	89.06	89.35	85.71	82.99	81.54	79.74
Models Gen. [98]	89.64	89.24	89.38	82.85	90.79	88.62	90.11	90.43	89.22	**85.21**	80.83	77.40
Swin UNETR [68]	89.56	90.80	93.08	86.38	**94.35**	89.65	**92.02**	**91.99**	**89.65**	82.20	85.01	**81.06**
UniMiSS [77]	89.20	91.21	**94.16**	86.61	91.57	87.29	90.18	90.56	88.09	83.47	80.73	76.40
Universal Model	88.95	**91.38**	93.82	**87.04**	93.53	88.96	90.50	91.40	89.18	84.25	**83.63**	79.95

Method	T5	T4	T3	T2	T1	C7	C6	C5	C4	C3	C2	C1	Average
Scratch	73.14	72.26	77.12	80.36	85.76	83.39	69.80	70.23	69.82	85.74	83.35	78.18	81.06
MedicalNet [10]	77.28	76.60	76.57	80.94	85.54	83.05	76.05	73.04	80.55	74.35	74.67	72.91	82.28
Models Gen. [98]	79.59	**78.73**	82.01	84.63	90.02	88.20	**81.09**	78.90	78.21	89.69	88.06	80.23	85.12
Swin UNETR [68]	82.33	77.74	81.78	83.53	88.22	87.81	78.38	80.36	83.00	92.68	87.97	80.16	86.23
UniMiSS [77]	78.97	76.60	82.33	85.14	90.04	88.68	79.18	79.17	79.00	88.19	86.38	79.80	85.12
Universal Model	**83.07**	78.67	**82.97**	**86.06**	**90.67**	**88.75**	77.03	**80.87**	**83.05**	**92.94**	**88.20**	**80.87**	**86.49**

Table 10. **The complete evaluation of TotalSeg_cardiac.** The results are evaluated by DSC. Our Universal Model represents the best transferability. The abbreviation in the table is listed as follows. HM (heart myocardium), HA (heart atrium), HV (heart ventricle), PA (pulmonary artery), IA (iliac artery), IV (iliac vena), UB (urinary bladder).

Method	esophagus	trachea	HM	HA_left	HV_left	HA_right	HV_right	PA	brain
Scratch	84.73	90.72	85.53	91.78	91.15	90.10	88.25	87.20	93.79
MedicalNet [10]	89.43	94.08	88.71	93.50	92.17	90.90	90.83	89.51	95.11
Models Gen. [98]	87.96	93.47	87.40	93.61	92.23	92.02	89.74	89.34	94.99
Swin UNETR [68]	89.77	94.37	88.85	94.42	92.99	92.61	90.40	88.91	95.14
UniMiSS [77]	90.45	94.51	90.29	94.34	93.70	93.10	91.46	89.67	94.99
Universal Model	**90.97**	**94.71**	**90.88**	**94.64**	**93.72**	**93.30**	**91.66**	**90.80**	**95.34**

Method	IA_left	IA_right	IV_left	IV_right	small_bow.	duodenum	colon	UB	face	Average
Scratch	80.32	79.78	79.80	81.69	81.97	72.21	82.51	89.59	69.40	84.47
MedicalNet [10]	87.06	84.90	86.93	86.46	83.14	72.01	84.22	90.43	73.85	87.40
Models Gen. [98]	85.71	83.09	85.77	85.79	81.75	69.37	85.25	90.31	69.42	86.51
Swin UNETR [68]	88.26	86.44	87.13	87.59	83.29	70.71	87.50	89.93	74.08	87.91
UniMiSS [77]	89.18	87.81	89.04	88.55	84.83	74.74	88.16	91.83	74.76	88.96
Universal Model	**89.89**	**88.54**	**89.58**	**89.27**	**84.85**	**76.23**	**89.06**	**92.07**	**76.81**	**89.57**

Table 11. **The complete evaluation of TotalSeg_muscles.** The results are evaluated by DSC. Our Universal Model represents the best transferability. The abbreviation in the table is listed as follows. Clav. (Clavicula), GMa (gluteus maximus), GMe (gluteus medius), GMi (gluteus minimus), Aotu. (Autochthon)

Method	Humerus_L	Humerus_R	Scapula_L	Scapula_R	Clav._L	Clav._R	Femur_L	Femur_R	Hip_L	Hip_R	Sacrum
Scratch	84.27	84.44	91.71	89.78	80.38	75.81	93.41	93.02	92.90	88.66	83.63
MedicalNet [10]	87.25	85.67	88.68	92.62	94.35	93.96	84.85	96.59	96.98	96.31	95.19
Models Gen. [98]	90.61	79.73	88.56	92.06	91.19	92.57	86.08	93.57	85.35	82.40	87.91
Swin UNETR [68]	88.32	86.35	90.82	93.88	94.90	94.52	85.92	97.71	97.42	97.49	95.73
UniMiSS [77]	89.73	92.30	91.72	94.77	94.57	93.66	84.92	97.67	97.35	97.11	96.18
Universal Model	**91.32**	**93.87**	**93.11**	**95.59**	**95.00**	**95.88**	**86.79**	**98.48**	**98.04**	**98.32**	**96.94**

Method	GMa_L	GMa_R	GMe_L	GMe_R	GMi_L	GMi_R	Aotu._L	Aotu._R	Iliopsoas_L	Iliopsoas_R	Average
Scratch	95.53	91.78	85.27	94.80	86.54	93.01	95.17	93.44	**87.99**	83.95	88.83
MedicalNet [10]	94.69	95.72	92.17	89.15	89.76	90.77	94.45	94.24	80.29	84.94	91.36
Models Gen. [98]	96.19	92.06	90.07	**94.99**	92.12	92.60	95.86	95.93	85.64	83.82	89.96
Swin UNETR [68]	95.32	96.34	93.57	89.87	90.75	91.74	95.16	94.86	83.53	86.00	92.39
UniMiSS [77]	95.53	96.37	93.80	90.28	90.87	93.02	95.17	95.48	85.71	84.02	92.86
Universal Model	**96.68**	**96.99**	**95.55**	91.36	**93.19**	**94.52**	**96.31**	**96.34**	86.92	**88.89**	**94.29**

Table 12. **The complete evaluation of TotalSeg_organs.** The results are evaluated by DSC. Our Universal Model represents the best transferability. The abbreviation in the table is listed as follows. IVC (inferior vena cava), PSV (portal vein and splenic vein), AG (adrenal gland), LUL (lung upper lobe), LLL (lung lower lobe), LML (lung middle lobe)

Method	spleen	Kidney_R	Kidney_L	gallbladder	liver	stomach	aorta	IVC	PSV
Scratch	93.58	94.09	87.73	73.86	96.79	89.17	90.68	82.10	71.35
MedicalNet [10]	95.54	92.43	90.86	79.36	97.10	91.53	90.12	86.18	73.34
Models Gen. [98]	95.60	94.37	88.51	78.39	97.39	91.68	93.18	85.94	74.58
Swin UNETR [68]	89.77	94.37	88.85	74.42	92.99	92.61	90.40	**88.91**	75.14
UniMiSS [77]	95.78	**94.75**	89.35	79.14	97.39	91.87	**93.50**	86.19	75.26
Universal Model	**96.24**	94.67	**91.43**	**81.48**	**97.63**	**92.76**	92.22	87.87	**76.10**

Method	pancreas	AG_R	AG_L	LUL_L	LLL_L	LUL_R	LML_R	LLL_R	Average
Scratch	80.80	78.94	72.83	95.88	91.66	87.17	88.91	93.71	86.42
MedicalNet [10]	83.11	79.15	69.22	93.64	89.88	86.38	87.08	92.40	86.90
Models Gen. [98]	82.97	**83.05**	**75.49**	95.79	92.90	90.10	91.06	94.65	85.78
Swin UNETR [68]	**85.24**	81.86	74.33	95.06	92.16	88.37	89.45	94.04	88.56
UniMiSS [77]	82.11	79.37	73.12	**96.08**	**93.18**	**90.31**	**91.99**	**95.43**	88.51
Universal Model	85.21	82.25	75.01	95.04	92.28	88.21	89.69	94.06	**88.95**

Table 13. **The complete evaluation of JHH.** The results are evaluated by DSC. IVC (inferior vena cava), PSV (portal vein and splenic vein), AG (adrenal gland), CAA (celiac abdominal aorta)

Method	spleen	Kidney_R	Kidney_L	gallbladder	liver	stomach
Scratch	95.66	94.43	93.69	86.14	96.74	94.30
MedicalNet [10]	91.08	88.63	86.60	61.23	93.29	88.22
Models Gen. [98]	95.02	93.44	93.07	84.73	94.12	94.05
Swin UNETR [68]	94.71	93.95	92.27	81.75	96.00	92.79
UniMiSS [77]	88.35	91.49	90.41	82.91	93.80	89.57
Universal Model	**95.98**	**94.71**	**94.00**	**87.18**	**96.87**	**94.50**

Method	aorta	IVC	pancreas	PSV	AG	CAA	Average
Scratch	87.68	79.73	85.03	68.48	66.61	50.61	81.98
MedicalNet [10]	83.27	75.32	70.67	46.82	41.69	26.87	68.88
Models Gen. [98]	**89.46**	**81.50**	84.23	**71.79**	**70.46**	**54.23**	**82.81**
Swin UNETR [68]	87.43	80.89	81.19	66.71	65.04	36.38	79.55
UniMiSS [77]	88.50	77.98	71.86	61.68	51.82	49.16	76.10
Universal Model	88.36	79.98	**85.82**	69.38	65.88	50.53	82.24

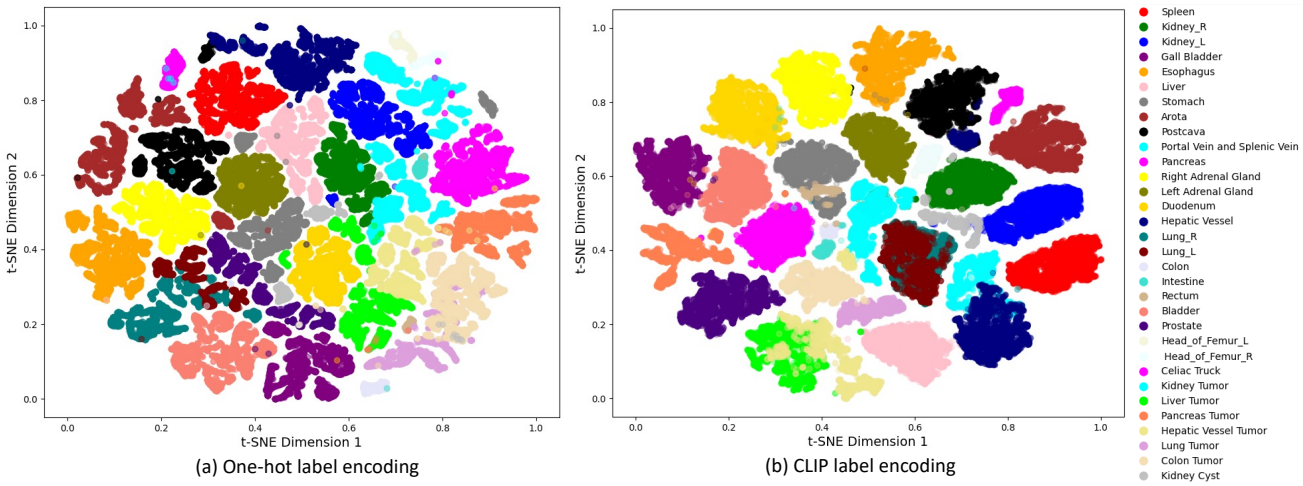

Figure 12. **t-SNE Visualization of Whole Embedding Space.** Colors for corresponding embeddings are shown in figure.

Table 14. **The complete results of embedding ablation study.** The results are evaluated by DSC. GB (Gallbladder), PSV (portal vein and splenic vein), AG (adrenal gland), HV (hepatic vessel), HF (head of femur), CT (celiac truck), KiT(kidney tumor), LiT (liver tumor), PT (pancreas tumor), HVT (hepatic vessel tumor), LuT (lung tumor), CoT (colon tumor), KiC (kideney cyst)

Embedding	spleen	Kidney_R	Kidney_L	GB	Esophagus	Liver	Stomach	Aorta	Postcava	PSV	Pancreas
One-hot [88]	91.92	91.98	92.14	71.75	70.28	95.10	80.52	83.57	82.71	67.81	74.06
BioBERT [83]	94.65	93.26	**92.98**	75.14	72.32	95.09	87.68	91.05	**83.91**	67.83	80.51
CLIP V1	92.35	91.83	91.89	72.45	71.38	90.23	73.07	86.77	78.17	74.00	74.91
CLIP V2	93.05	92.14	91.42	**75.88**	**75.56**	94.75	75.79	91.15	80.64	**78.90**	78.94
CLIP V3	**94.69**	**94.09**	92.77	73.45	72.87	**95.71**	**89.19**	**92.19**	83.44	59.20	**86.09**

Embedding	AG_R	AG_L	Duodenum	HV	Lung_R	Lung_L	Colon	Intestine	Rectum	Bladder	Prostate
One-hot [88]	64.52	66.96	55.66	71.03	**79.63**	66.75	69.22	78.05	69.87	76.74	66.15
BioBERT [83]	65.94	68.72	**68.61**	59.14	75.40	69.09	71.24	**81.78**	65.58	74.51	69.51
CLIP V1	72.07	72.42	62.42	**74.53**	79.32	76.52	70.32	75.65	63.11	75.06	66.47
CLIP V2	**79.98**	**79.73**	66.01	68.65	75.87	**82.98**	**74.88**	70.82	64.64	70.06	68.8
CLIP V3	64.75	70.18	71.11	65.43	77.48	62.11	71.77	81.47	**79.42**	**86.71**	**72.96**

Embedding	HF_L	HF_R	CT	KiT	LiT	PT	HVT	LuT	CoT	KiC	Ave
One-hot [88]	70.27	60.23	78.92	63.84	68.02	55.48	52.31	53.87	48.39	**35.81**	70.42
BioBERT [83]	74.39	79.07	80.69	57.41	63.44	39.70	57.88	58.57	54.19	20.33	71.55
CLIP V1	74.61	72.53	79.28	56.62	76.24	61.05	56.49	73.60	55.03	32.87	73.49
CLIP V2	69.98	75.73	**84.04**	67.04	**82.09**	**77.75**	67.45	**75.38**	55.55	35.79	75.66
CLIP V3	**84.94**	**89.45**	77.55	**68.72**	74.87	65.46	**73.53**	73.12	**60.66**	30.44	**76.11**

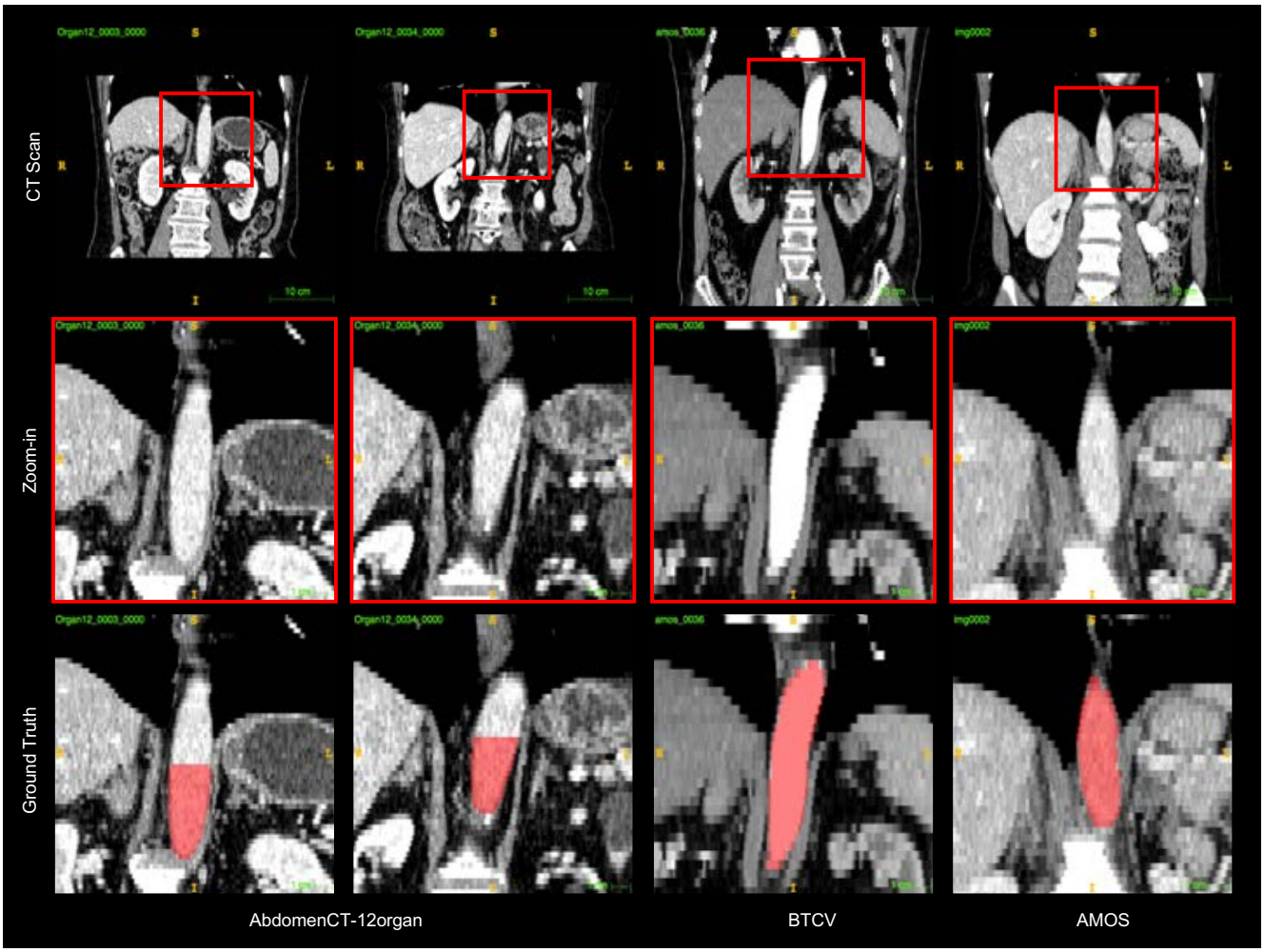

Figure 13. **Inconsistent Label Protocol.** The aorta annotation standard is inconsistent in AbdomenCT-12organ and other datasets. A part of the upper aorta region is missing in AbdomenCT-12organ, while the aorta annotation is complete in BTCV and AMOS.

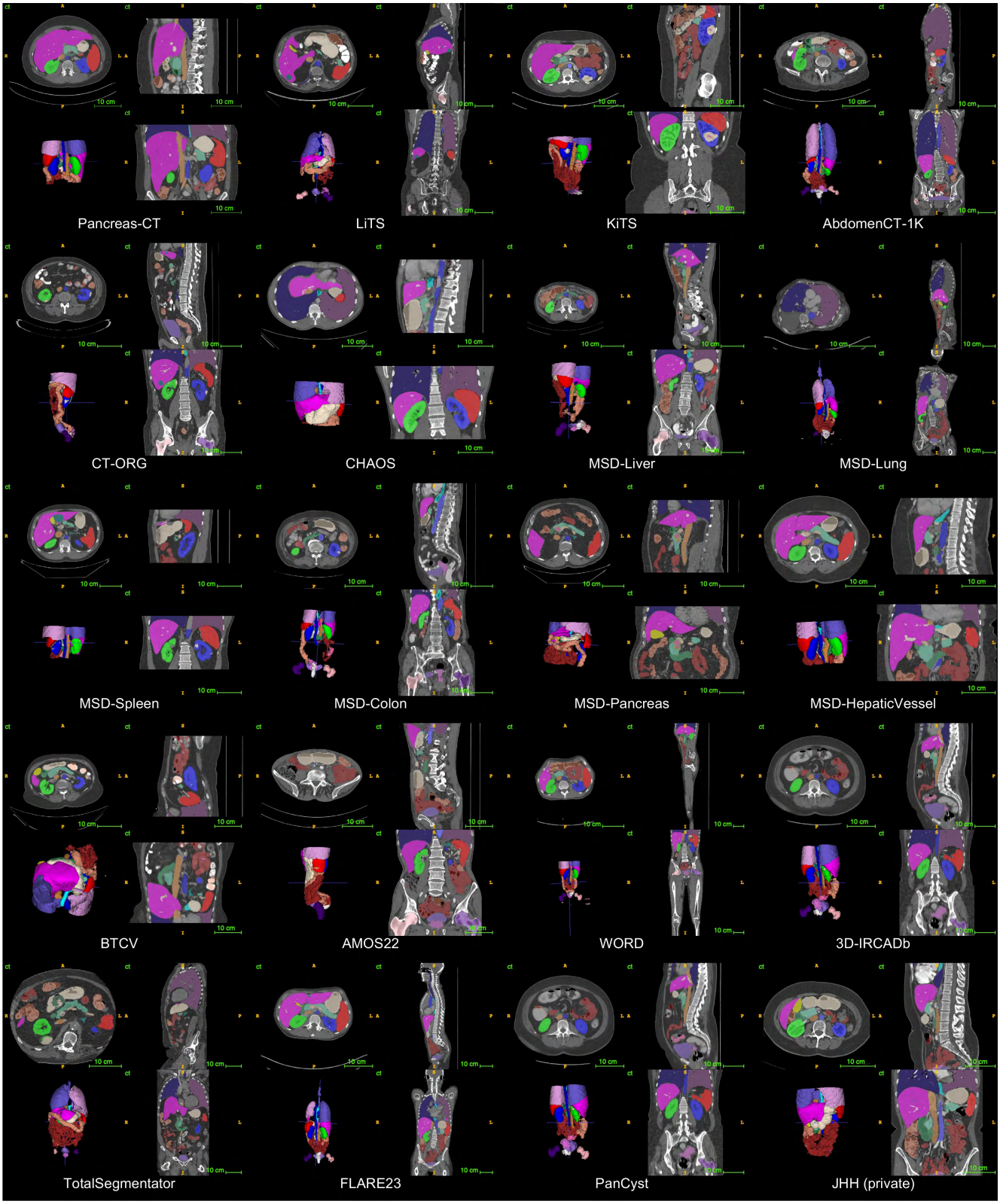

Figure 14. **Prediction of incomplete labels in previous datasets.** We leveraged the predictions generated by the Universal Model to produce masks for 25 organs in 20 CT datasets, achieving a satisfactory level of accuracy. However, we note that the accuracy of the 6-tumor segmentation still requires validation through pathology reports, which we have identified as a future direction for our work.

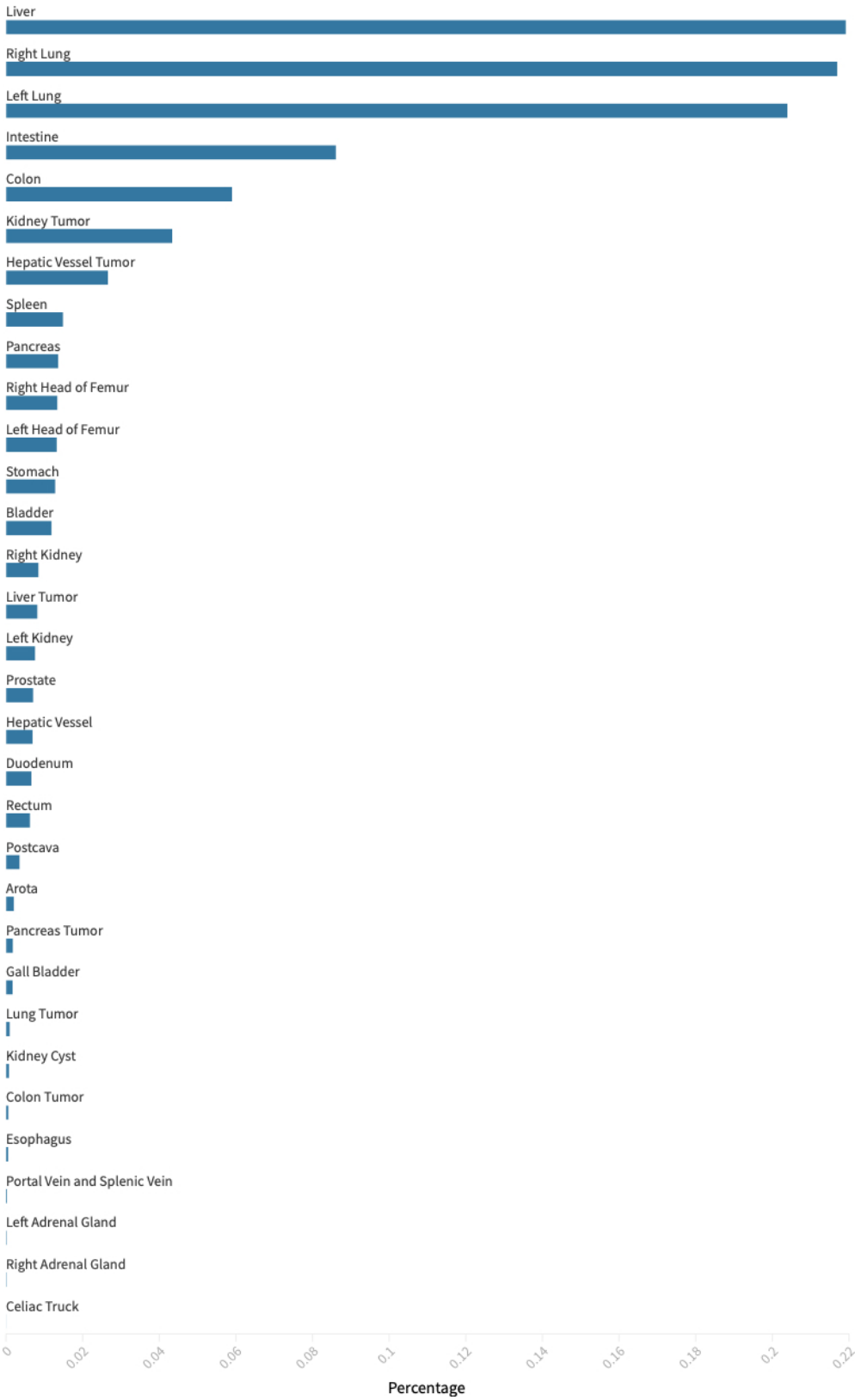

Figure 15. **The proportion of 32 classes.** We observe that the assembly of datasets presents severe long-tail distribution.