CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection

Jie Liu¹, Yixiao Zhang², Jie-Neng Chen², Junfei Xiao², Yongyi Lu², Bennett A. Landman³, Yixuan Yuan⁴,⁵, Alan Yuille², Yucheng Tang³,⁶,*, and Zongwei Zhou²,*

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

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
The Universal Model can be generalized to CT scans from and detection of 6 tumors with state-of-the-art performance. That allows superior segmentation of 25 organs exploiting the semantic relationship between images and language.

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 BTVC, 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 BTVC. (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. 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 BTVC 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, 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

\[\text{CLIP (Contrastive Language–Image Pre-training) was pre-trained on 400 million image-text pairs (some are medical images and text [5]), exploiting the semantic relationship between images and language.}\]
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 numerous pivotal 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 $D_i = \{ (X_i, Y_i) \}_{n=1}^{N}$, 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 $X_i$ is annotated in $Y_i$, $D$ is a fully labeled dataset; otherwise, $D$ is a partially labeled dataset.

**Previous solutions.** Two groups of solutions were proposed to address the partial label problem. Given a data point $X_n, n \in [1, N]$, the objective is to train a model $F(\cdot)$ using the assembly dataset $D_A = \{ D_1, D_2, ..., D_M \}$, and the model can predict all $K$ classes, if presented in $X_n$.

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

- **Solution #2** [88, 32, 99] aims to solve $F_{\theta}(X_n, w_k) = P_n, n \in [1, N], k \in [1, K]$, where $w_k$ 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
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 $w_k$ in Solution #2 to CLIP embedding and introduce 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.

**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 name). 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 [69], to generate parameters ($\theta_k$), i.e., $\theta_k = \text{MLP}(w_k \oplus f)$, 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. The standardized and normalized CT scans are then processed by the vision encoder. Let $F$ be the image features extracted by the vision encoder. To process $F$, 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 [CLS]$_k$. The prediction for the class [CLS]$_k$ is computed as $P_k = \text{Sigmoid}((F \ast \theta_{k_1}) \ast \theta_{k_2} \ast \theta_{k_3})$, where $\theta_k = \{\theta_{k_1}, \theta_{k_2}, \theta_{k_3}\}$ are computed in the text branch, and $*$ represents the convolution. For each class [CLS]$_k$, we generate the prediction $P_k \in \mathbb{R}^{1 \times D \times W \times H}$ 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 $Y'$ 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. 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 obtain 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-
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].

<table>
<thead>
<tr>
<th>Method</th>
<th>Task03 Liver</th>
<th>Task07 Pancreas</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dice1</td>
<td>Dice2</td>
</tr>
<tr>
<td>Kim et al. [33]</td>
<td>94.25</td>
<td>72.96</td>
</tr>
<tr>
<td>Trans VW [21]</td>
<td>95.18</td>
<td>76.90</td>
</tr>
<tr>
<td>C2FNAS [66]</td>
<td>94.98</td>
<td>72.89</td>
</tr>
<tr>
<td>Models Gen. [97]</td>
<td>95.72</td>
<td>77.50</td>
</tr>
<tr>
<td>nnUNet [29]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DiNTS [23]</td>
<td>95.35</td>
<td>74.62</td>
</tr>
<tr>
<td>Swin UNETR [68]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Universal Model</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Task08 Hepatic Vessel</th>
<th>Task06 Lung</th>
<th>Task09 Spleen</th>
<th>Task10 Colon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dice1</td>
<td>Dice2</td>
<td>Avg.</td>
<td>NSD1</td>
</tr>
<tr>
<td>Kim et al. [33]</td>
<td>62.34</td>
<td>68.63</td>
<td>65.49</td>
<td>83.22</td>
</tr>
<tr>
<td>Trans VW [21]</td>
<td>65.80</td>
<td>71.44</td>
<td>69.62</td>
<td>84.01</td>
</tr>
<tr>
<td>C2FNAS [86]</td>
<td>64.30</td>
<td>71.00</td>
<td>67.65</td>
<td>83.78</td>
</tr>
<tr>
<td>Models Gen. [97]</td>
<td>65.80</td>
<td>71.44</td>
<td>68.62</td>
<td>84.01</td>
</tr>
<tr>
<td>nnUNet [29]</td>
<td>66.46</td>
<td>71.78</td>
<td>69.12</td>
<td>84.43</td>
</tr>
<tr>
<td>DiNTS [23]</td>
<td>64.50</td>
<td>71.76</td>
<td>68.13</td>
<td>83.98</td>
</tr>
<tr>
<td>Swin UNETR [68]</td>
<td>65.69</td>
<td>72.20</td>
<td>68.95</td>
<td>84.83</td>
</tr>
<tr>
<td>Universal Model</td>
<td>67.15</td>
<td>75.86</td>
<td>71.51</td>
<td>84.84</td>
</tr>
</tbody>
</table>

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.

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 independent 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. 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, pancreatic, and colon tumors. The quality of 19 other organs and 6 tumors has not been compared with human annotations because there is no publicly available CT scans that 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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Spl</th>
<th>RKid</th>
<th>LKid</th>
<th>Gall</th>
<th>Eso</th>
<th>Liv</th>
<th>Sto</th>
<th>Aor</th>
<th>IVC</th>
<th>Veins</th>
<th>Pan</th>
<th>AG</th>
</tr>
</thead>
<tbody>
<tr>
<td>RandPatch [67]</td>
<td>95.82</td>
<td>88.52</td>
<td>90.14</td>
<td>68.31</td>
<td>75.01</td>
<td>96.48</td>
<td>82.93</td>
<td>89.86</td>
<td>82.49</td>
<td>73.54</td>
<td>75.48</td>
<td>66.09</td>
</tr>
<tr>
<td>TransBTS [29]</td>
<td>94.59</td>
<td>89.23</td>
<td>90.47</td>
<td>68.50</td>
<td>75.59</td>
<td>96.14</td>
<td>83.72</td>
<td>88.85</td>
<td>82.28</td>
<td>74.25</td>
<td>75.12</td>
<td>66.74</td>
</tr>
<tr>
<td>nnFormer [91]</td>
<td>94.51</td>
<td>88.49</td>
<td>93.39</td>
<td>65.51</td>
<td>74.49</td>
<td>96.10</td>
<td>83.83</td>
<td>88.91</td>
<td>80.58</td>
<td>75.94</td>
<td>77.71</td>
<td>68.19</td>
</tr>
<tr>
<td>UNETR [22]</td>
<td>94.91</td>
<td>92.10</td>
<td>93.12</td>
<td>76.98</td>
<td>74.14</td>
<td>96.17</td>
<td>79.98</td>
<td>89.74</td>
<td>81.20</td>
<td>75.05</td>
<td>80.12</td>
<td>62.60</td>
</tr>
<tr>
<td>nnUNet [29]</td>
<td><strong>95.92</strong></td>
<td>88.28</td>
<td>92.62</td>
<td>66.58</td>
<td>75.71</td>
<td>96.49</td>
<td>86.05</td>
<td>83.33</td>
<td>82.72</td>
<td><strong>78.31</strong></td>
<td>79.17</td>
<td>67.99</td>
</tr>
<tr>
<td>Swin UNETR [68]</td>
<td>95.44</td>
<td>93.38</td>
<td>93.40</td>
<td>77.12</td>
<td>74.14</td>
<td>96.39</td>
<td>80.12</td>
<td>82.93</td>
<td>75.08</td>
<td>81.02</td>
<td>64.98</td>
<td>82.06</td>
</tr>
</tbody>
</table>

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 highly 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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Liver Tumor</th>
<th>Kidney Tumor</th>
<th>Pancreatic Tumor</th>
</tr>
</thead>
<tbody>
<tr>
<td>nnU-Net [29]</td>
<td>94.44</td>
<td>75.00</td>
<td>83.60</td>
</tr>
<tr>
<td>UNet++ [96]</td>
<td>94.44</td>
<td>80.00</td>
<td>86.62</td>
</tr>
<tr>
<td>UNETR [22]</td>
<td>86.11</td>
<td>95.00</td>
<td>90.34</td>
</tr>
<tr>
<td>Swin UNETR [68]</td>
<td>91.67</td>
<td>85.00</td>
<td>88.21</td>
</tr>
<tr>
<td>Universal Model</td>
<td>88.89</td>
<td>95.00</td>
<td>91.84</td>
</tr>
</tbody>
</table>

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.

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.

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 conventional one-hot labels (DoDNet [88]) and text-only pre-trained embedding (BioLinkBERT [83]).

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

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

9Existing dataset-specific models are limited to being trained individually 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>TotalSeg_vertebrae</th>
<th>TotalSeg_cardiac</th>
<th>TotalSeg_muscles</th>
<th>TotalSeg_organ</th>
<th>JHH_cardiac</th>
<th>JHH_organ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scratch</td>
<td>81.06</td>
<td>84.47</td>
<td>88.83</td>
<td>86.42</td>
<td>71.63</td>
<td>89.08</td>
</tr>
<tr>
<td>MedicalNet [10]</td>
<td>82.28</td>
<td>87.40</td>
<td>91.36</td>
<td>86.90</td>
<td>58.07</td>
<td>77.68</td>
</tr>
<tr>
<td>Models Gen. [98]</td>
<td>85.12</td>
<td>86.51</td>
<td>89.96</td>
<td>85.78</td>
<td>74.25</td>
<td>88.64</td>
</tr>
<tr>
<td>Swin UNETR [68]</td>
<td>86.23</td>
<td>87.91</td>
<td>92.39</td>
<td>88.56</td>
<td>67.85</td>
<td>87.21</td>
</tr>
<tr>
<td>UnMiS [77]</td>
<td>85.12</td>
<td>88.96</td>
<td>92.86</td>
<td>88.51</td>
<td>69.33</td>
<td>82.53</td>
</tr>
<tr>
<td>Universal Model</td>
<td>86.49</td>
<td>89.57</td>
<td>94.43</td>
<td>88.95</td>
<td>72.06</td>
<td>89.37</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>TotalSeg_vertebrae</th>
<th>TotalSeg_cardiac</th>
<th>TotalSeg_muscles</th>
<th>TotalSeg_organ</th>
<th>JHH_cardiac</th>
<th>JHH_organ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scratch</td>
<td>81.06</td>
<td>84.47</td>
<td>88.83</td>
<td>86.42</td>
<td>71.63</td>
<td>89.08</td>
</tr>
<tr>
<td>MedicalNet [10]</td>
<td>82.28</td>
<td>87.40</td>
<td>91.36</td>
<td>86.90</td>
<td>58.07</td>
<td>77.68</td>
</tr>
<tr>
<td>Models Gen. [98]</td>
<td>85.12</td>
<td>86.51</td>
<td>89.96</td>
<td>85.78</td>
<td>74.25</td>
<td>88.64</td>
</tr>
<tr>
<td>Swin UNETR [68]</td>
<td>86.23</td>
<td>87.91</td>
<td>92.39</td>
<td>88.56</td>
<td>67.85</td>
<td>87.21</td>
</tr>
<tr>
<td>UnMiS [77]</td>
<td>85.12</td>
<td>88.96</td>
<td>92.86</td>
<td>88.51</td>
<td>69.33</td>
<td>82.53</td>
</tr>
<tr>
<td>Universal Model</td>
<td>86.49</td>
<td>89.57</td>
<td>94.43</td>
<td>88.95</td>
<td>72.06</td>
<td>89.37</td>
</tr>
</tbody>
</table>

**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%, 88.95%, 94.43% and 88.95% for four downstream tasks in the TotalSegmentator dataset (see Table 6).

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.


[38] Xiaokun Liang, Na Li, Zhicheng Zhan, 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.


[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.


[70] Constantin Ulrich, Fabian Isensee, Tassilo Wald, Maximilian Zenk, Michael Baumgartner, and Klaus H Maier-Hein. 11


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.

<table>
<thead>
<tr>
<th>Datasets</th>
<th># Targets</th>
<th># Scans</th>
<th>Annotated Organs or Tumors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Pancreas-CT [60]</td>
<td>1</td>
<td>82</td>
<td>Pancreas</td>
</tr>
<tr>
<td>2. LiTS [3]</td>
<td>2</td>
<td>201</td>
<td>Liver, Liver Tumor*</td>
</tr>
<tr>
<td>3. KiTS [23]</td>
<td>2</td>
<td>300</td>
<td>Kidney, Kidney Tumor*</td>
</tr>
<tr>
<td>4. Abdomen-CT1K [43]</td>
<td>4</td>
<td>1,000</td>
<td>Spleen, Kidney, Liver, Pancreas</td>
</tr>
<tr>
<td>5. CT-ORG [58]</td>
<td>4</td>
<td>140</td>
<td>Lung, Liver, Kidneys and Bladder</td>
</tr>
<tr>
<td>6. CHAOS [71]</td>
<td>4</td>
<td>40</td>
<td>Liver, Left Kidney, Right Kidney, Spl</td>
</tr>
<tr>
<td>7-11. MSD CT Tasks [1]</td>
<td>9</td>
<td>947</td>
<td>Spl, Liver and Tumor*, Lung Tumor*, Colon Tumor*, Pan and Tumor*, Hepatic Vessel and Tumor*</td>
</tr>
<tr>
<td>12. BTCV [56]</td>
<td>13</td>
<td>50</td>
<td>Spl, RKid, LKid, Gall, Eso, Liv, Sto, Aor, IVC, R&amp;SVeins, Pan, RAG, LAG</td>
</tr>
<tr>
<td>13. AMOS22 [33]</td>
<td>15</td>
<td>500</td>
<td>Spl, RKid, LKid, Gall, Eso, Liv, Sto, Aor, IVC, Pan, RAG, LAG, Duo, Bla, ProUTE</td>
</tr>
<tr>
<td>14. WORD [42]</td>
<td>16</td>
<td>150</td>
<td>Spl, RKid, LKid, Gall, Eso, Liv, Sto, Pan, RAG, Duo, Col, Int, Rec, Bla, LFH, RFH</td>
</tr>
<tr>
<td>15. 3D-IRCADb [65]</td>
<td>13</td>
<td>20</td>
<td>Liv, Liv Cyst, RLang, LLang, Venous, PVein, Aor, Spl, RKid, LKid, Gall, IVC</td>
</tr>
<tr>
<td>16. TotalSegmentator [76]</td>
<td>104</td>
<td>1,024</td>
<td>Clavicula, Humerus, Scapula, Rib 1-12, Vertebræ C1-7, Vertebræ T1-9, Vertebræ 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, Aor, Spl, Liv, Gall, Pan, Kid, Eso, Sto, Duo, Small bowel, Colon, Bla, Autochthon, Iliopsoas, Gluteus Minimus, Gluteus Medius, Gluteus Maximus</td>
</tr>
<tr>
<td>17. JHH (private)</td>
<td>21</td>
<td>5,038</td>
<td>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*</td>
</tr>
</tbody>
</table>

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 MONAI10. The orientation of CT scans is changed into the specified axcodes. Isotropic spacing is adopted to re-slice each scan to the same voxel size of $1.5 \times 1.5 \times 1.5mm^3$. We truncate the intensity in each scan to the range $[-175, 250]$.

10https://monai.io/

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 × 96 × 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.

C.2. Network Structures

Text branch. We apply the pre-trained text encoder “ViT-B/32” of the CLIP as the text branch11. 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.

11https://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 $H_2 \times W_2 \times D_2$, 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:

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

(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.

<table>
<thead>
<tr>
<th>Method</th>
<th>L5</th>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>T2</th>
<th>T1</th>
<th>T11</th>
<th>T10</th>
<th>T9</th>
<th>T8</th>
<th>T7</th>
<th>T6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scratch</td>
<td>86.68</td>
<td>88.37</td>
<td>89.83</td>
<td>84.28</td>
<td>91.98</td>
<td>87.45</td>
<td>88.29</td>
<td>86.78</td>
<td>83.50</td>
<td>75.70</td>
<td>77.73</td>
<td>75.84</td>
<td></td>
</tr>
<tr>
<td>MedicalNet [10]</td>
<td><strong>91.72</strong></td>
<td>91.01</td>
<td>86.03</td>
<td>84.73</td>
<td>91.52</td>
<td><strong>89.98</strong></td>
<td>89.06</td>
<td>89.35</td>
<td>85.71</td>
<td>82.99</td>
<td>81.54</td>
<td>79.74</td>
<td></td>
</tr>
<tr>
<td>Models Gen. [98]</td>
<td>89.64</td>
<td>89.24</td>
<td>89.38</td>
<td>82.85</td>
<td>90.79</td>
<td>88.62</td>
<td>90.11</td>
<td>90.43</td>
<td>89.22</td>
<td><strong>85.21</strong></td>
<td>80.83</td>
<td>77.40</td>
<td></td>
</tr>
<tr>
<td>SwinUNETR [68]</td>
<td>89.56</td>
<td>90.80</td>
<td>93.08</td>
<td>86.38</td>
<td><strong>94.35</strong></td>
<td>89.65</td>
<td><strong>92.02</strong></td>
<td><strong>91.99</strong></td>
<td><strong>89.65</strong></td>
<td>82.20</td>
<td>85.01</td>
<td><strong>81.06</strong></td>
<td></td>
</tr>
<tr>
<td>UniMiSS [77]</td>
<td>89.20</td>
<td>91.21</td>
<td><strong>94.16</strong></td>
<td>86.61</td>
<td>91.57</td>
<td>87.29</td>
<td>90.18</td>
<td>90.56</td>
<td>88.09</td>
<td>83.47</td>
<td>80.73</td>
<td>76.40</td>
<td></td>
</tr>
<tr>
<td>Universal Model</td>
<td>88.95</td>
<td><strong>91.38</strong></td>
<td>93.82</td>
<td><strong>87.04</strong></td>
<td>93.53</td>
<td>88.96</td>
<td>90.50</td>
<td>90.40</td>
<td>89.18</td>
<td>84.25</td>
<td><strong>83.63</strong></td>
<td>79.95</td>
<td></td>
</tr>
</tbody>
</table>

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).

<table>
<thead>
<tr>
<th>Method</th>
<th>esophagus</th>
<th>trachea</th>
<th>HM</th>
<th>HA_left</th>
<th>HV_left</th>
<th>HA_right</th>
<th>HV_right</th>
<th>PA</th>
<th>brain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scratch</td>
<td>84.73</td>
<td>90.72</td>
<td>85.53</td>
<td>91.78</td>
<td>91.15</td>
<td>90.10</td>
<td>88.25</td>
<td>87.20</td>
<td>93.79</td>
</tr>
<tr>
<td>MedicalNet [10]</td>
<td>89.43</td>
<td>94.08</td>
<td>88.71</td>
<td>93.50</td>
<td>92.17</td>
<td>90.90</td>
<td>90.83</td>
<td>89.51</td>
<td>95.11</td>
</tr>
<tr>
<td>Models Gen. [98]</td>
<td>87.96</td>
<td>93.47</td>
<td>87.40</td>
<td>93.61</td>
<td>92.23</td>
<td>92.02</td>
<td>89.74</td>
<td>89.34</td>
<td>94.99</td>
</tr>
<tr>
<td>SwinUNETR [68]</td>
<td>87.99</td>
<td>94.37</td>
<td>88.85</td>
<td>94.42</td>
<td>92.99</td>
<td>92.61</td>
<td>90.40</td>
<td>88.91</td>
<td>95.14</td>
</tr>
<tr>
<td>UniMiSS [77]</td>
<td>90.45</td>
<td>94.51</td>
<td>90.29</td>
<td>94.34</td>
<td>93.70</td>
<td>93.10</td>
<td>91.46</td>
<td>89.67</td>
<td>94.99</td>
</tr>
<tr>
<td>Universal Model</td>
<td><strong>90.97</strong></td>
<td><strong>94.71</strong></td>
<td><strong>90.88</strong></td>
<td><strong>94.64</strong></td>
<td><strong>93.72</strong></td>
<td><strong>93.30</strong></td>
<td><strong>91.66</strong></td>
<td><strong>90.80</strong></td>
<td><strong>95.34</strong></td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th>Method</th>
<th>Humerus_L</th>
<th>Humerus_R</th>
<th>Scapula_L</th>
<th>Scapula_R</th>
<th>Clav_L</th>
<th>Clav_R</th>
<th>Femur_L</th>
<th>Femur_R</th>
<th>Hip_L</th>
<th>Hip_R</th>
<th>Sacrum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scratch</td>
<td>84.27</td>
<td>84.44</td>
<td>91.71</td>
<td>89.78</td>
<td>80.38</td>
<td>75.81</td>
<td>93.41</td>
<td>93.02</td>
<td>92.90</td>
<td>88.66</td>
<td>83.63</td>
</tr>
<tr>
<td>MedicalNet [10]</td>
<td>87.25</td>
<td>85.67</td>
<td>88.68</td>
<td>92.62</td>
<td>94.35</td>
<td>93.96</td>
<td>84.85</td>
<td>86.59</td>
<td>96.98</td>
<td>96.31</td>
<td>95.19</td>
</tr>
<tr>
<td>Models Gen. [98]</td>
<td>90.61</td>
<td>79.73</td>
<td>88.56</td>
<td>92.06</td>
<td>91.19</td>
<td>92.57</td>
<td>86.08</td>
<td>93.57</td>
<td>85.55</td>
<td>82.40</td>
<td>87.91</td>
</tr>
<tr>
<td>SwinUNETR [68]</td>
<td>88.32</td>
<td>86.35</td>
<td>90.82</td>
<td>93.88</td>
<td>94.90</td>
<td>94.52</td>
<td>85.92</td>
<td>97.57</td>
<td>97.42</td>
<td>97.49</td>
<td>95.73</td>
</tr>
<tr>
<td>UniMiSS [77]</td>
<td>89.73</td>
<td>92.30</td>
<td>91.72</td>
<td>94.77</td>
<td>94.57</td>
<td>93.66</td>
<td>84.92</td>
<td>97.67</td>
<td>97.35</td>
<td>97.11</td>
<td>96.18</td>
</tr>
<tr>
<td>Universal Model</td>
<td><strong>91.32</strong></td>
<td><strong>93.87</strong></td>
<td><strong>93.11</strong></td>
<td><strong>95.59</strong></td>
<td><strong>95.00</strong></td>
<td><strong>93.19</strong></td>
<td><strong>94.52</strong></td>
<td><strong>96.31</strong></td>
<td><strong>96.34</strong></td>
<td><strong>86.92</strong></td>
<td><strong>88.89</strong></td>
</tr>
</tbody>
</table>

---

19
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)

<table>
<thead>
<tr>
<th>Method</th>
<th>spleen</th>
<th>Kidney_R</th>
<th>Kidney_L</th>
<th>gallbladder</th>
<th>liver</th>
<th>stomach</th>
<th>aorta</th>
<th>IVC</th>
<th>PSV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scratch</td>
<td>93.58</td>
<td>94.09</td>
<td>87.73</td>
<td>73.86</td>
<td>96.79</td>
<td>89.17</td>
<td>90.68</td>
<td>82.10</td>
<td>71.35</td>
</tr>
<tr>
<td>MedicalNet [10]</td>
<td>95.54</td>
<td>92.43</td>
<td>90.86</td>
<td>79.36</td>
<td>97.10</td>
<td>91.53</td>
<td>90.12</td>
<td>86.18</td>
<td>73.34</td>
</tr>
<tr>
<td>Models Gen. [98]</td>
<td>95.60</td>
<td>94.37</td>
<td>88.51</td>
<td>78.39</td>
<td>97.39</td>
<td>91.68</td>
<td>93.18</td>
<td>85.94</td>
<td>74.58</td>
</tr>
<tr>
<td>Swin UNETR [68]</td>
<td>89.77</td>
<td>94.37</td>
<td>88.85</td>
<td>74.42</td>
<td>92.99</td>
<td>92.61</td>
<td>90.40</td>
<td>88.91</td>
<td>75.14</td>
</tr>
<tr>
<td>UniMiSS [77]</td>
<td>95.76</td>
<td>94.75</td>
<td>89.35</td>
<td>79.14</td>
<td>97.39</td>
<td>91.87</td>
<td>93.50</td>
<td>86.19</td>
<td>75.26</td>
</tr>
<tr>
<td>Universal Model</td>
<td>96.24</td>
<td>94.67</td>
<td>91.43</td>
<td>81.48</td>
<td>97.63</td>
<td>92.76</td>
<td>92.22</td>
<td>87.87</td>
<td>76.10</td>
</tr>
</tbody>
</table>

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

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)

<table>
<thead>
<tr>
<th>Method</th>
<th>spleen</th>
<th>Kidney_R</th>
<th>Kidney_L</th>
<th>gallbladder</th>
<th>liver</th>
<th>stomach</th>
<th>aorta</th>
<th>IVC</th>
<th>PSV</th>
<th>AG</th>
<th>CAA</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scratch</td>
<td>95.66</td>
<td>94.43</td>
<td>93.69</td>
<td>86.14</td>
<td>96.74</td>
<td>94.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MedicalNet [10]</td>
<td>91.08</td>
<td>88.63</td>
<td>86.60</td>
<td>61.23</td>
<td>93.29</td>
<td>88.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Models Gen. [98]</td>
<td>95.02</td>
<td>93.44</td>
<td>93.07</td>
<td>84.73</td>
<td>94.12</td>
<td>94.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swin UNETR [68]</td>
<td>94.71</td>
<td>93.95</td>
<td>92.27</td>
<td>81.75</td>
<td>96.00</td>
<td>92.79</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UniMiSS [77]</td>
<td>88.35</td>
<td>91.49</td>
<td>90.41</td>
<td>82.91</td>
<td>93.80</td>
<td>89.57</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Universal Model</td>
<td>95.98</td>
<td>94.71</td>
<td>94.00</td>
<td>87.18</td>
<td>96.87</td>
<td>94.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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 (kidney cyst)

<table>
<thead>
<tr>
<th>Embedding</th>
<th>spleen</th>
<th>Kidney_R</th>
<th>Kidney_L</th>
<th>GB</th>
<th>Esophagus</th>
<th>Liver</th>
<th>Stomach</th>
<th>Aorta</th>
<th>Postcava</th>
<th>PSV</th>
<th>Pancreas</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-hot [88]</td>
<td>91.92</td>
<td>91.98</td>
<td>92.14</td>
<td>71.75</td>
<td>70.28</td>
<td>95.10</td>
<td>80.52</td>
<td>83.57</td>
<td>82.71</td>
<td>67.81</td>
<td>74.06</td>
</tr>
<tr>
<td>BioBERT [83]</td>
<td>94.65</td>
<td>93.26</td>
<td>92.98</td>
<td>75.14</td>
<td>72.32</td>
<td>95.09</td>
<td>87.68</td>
<td>91.05</td>
<td>83.91</td>
<td>67.83</td>
<td>80.51</td>
</tr>
<tr>
<td>CLIP V1</td>
<td>92.35</td>
<td>91.83</td>
<td>91.89</td>
<td>72.45</td>
<td>71.38</td>
<td>90.23</td>
<td>73.07</td>
<td>86.77</td>
<td>78.17</td>
<td>74.00</td>
<td>74.91</td>
</tr>
<tr>
<td>CLIP V2</td>
<td>93.05</td>
<td>92.14</td>
<td>91.42</td>
<td>75.88</td>
<td>75.56</td>
<td>94.75</td>
<td>75.79</td>
<td>91.15</td>
<td>80.64</td>
<td>78.90</td>
<td>78.94</td>
</tr>
<tr>
<td>CLIP V3</td>
<td>94.69</td>
<td>94.09</td>
<td>92.77</td>
<td>75.45</td>
<td>72.87</td>
<td>95.71</td>
<td>89.19</td>
<td>92.19</td>
<td>84.44</td>
<td>79.20</td>
<td>86.09</td>
</tr>
</tbody>
</table>

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

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