CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection

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

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

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 seg-

\(^*\)Corresponding authors: Yucheng Tang (yuchengt@nvidia.com) and Zongwei Zhou (zzhou82@jh.edu)
downstream tasks) [97]. Specifically, we have assembled 14 publicly available datasets, including 3,410 CT scans with 25 partially annotated organs and 6 tumors. Formidable challenges exist in assembling partially annotated datasets. First, label inconsistency, in five aspects. (i) Index inconsistency. The same organ can be labeled as different indexes. For example, the stomach is labeled ‘7’ in BTCV, but ‘5’ in WORD. (ii) Name inconsistency. Naming can be confusing if multiple labels refer to the same anatomical structure. For example, “postcava” in AMOS22 and “inferior vena cava” in BTCV. (iii) Background inconsistency. For example, when combining Pancreas-CT and MSD-Spleen, the pancreas is marked as the background in MSD-Spleen, but it should have been marked as the foreground. (iv) Organ overlapping. There is overlap between various organs. For example, “Hepatic Vessel” is part of the “Liver” and “Kidney Tumor” is a sub-volume of the “Kidney”. (v) Data overlapping. Some CT scans are overlapped among public datasets, but with different annotations. For example, KiTS is part of AbdomenCT-1K, and kidney tumor is annotated in KiTS rather than AbdomenCT-1K. Second, label orthogonality. Most segmentation methods, trained with one-hot labels [88], ignore the semantic relationship between classes. Given one-hot labels of liver [1,0,0], liver tumor [0,1,0], and pancreas [0,0,1], there is no semantic difference between liver→liver tumor and liver→pancreas. A possible solution is few-hot labels [62], with which, the liver, liver tumor, and pancreas can be encoded as [1,0,0], [1,1,0], and [0,0,1]. Although few-hot labels could indicate that liver tumors are part of the liver, the relationship between organs remains orthogonal.

To address above mentioned challenges, 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 BTCV challenges, leading to substantial performance improvement. Moreover, six organs can be annotated by Universal Model with a similar intra-observer variability to humans.

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

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

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

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

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

2. Related Work

Partial label problem. Publicly available datasets for abdominal imaging focus on different organs and tumors [35, 43, 42, 31], e.g., AbdomenCT-1K dataset for 4 organ segmentation [43], WORD dataset for 16 organ segmentation [42] and TotalSegmentor dataset for 104 anatomical structure segmentation [76]. The partial label problem occurs when training AI models on a combination of these datasets due to their inconsistent label taxonomy. To exploit the partial labels, several approaches have been investigated [93, 18, 88, 89], aiming for a single model that can perform organ segmentation [39, 12] and tumor detection [2, 100, 80, 40, 48, 78, 45]. These studies have the following limitations. (1) Due to the small scale of the dataset assembly2, the potential of assembling datasets was not convincing. Their performance was similar to dataset-specific models and was not evaluated on the official benchmark. (2) Due to the one-hot labels, the semantic relationship between organs and tumors was discarded. Table 1 reveals that the introduction of CLIP embedding is a salient factor to our proposed framework.

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

---

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

2 Zhou et al. [93] assembled 150 CT scans from 4 datasets; Fang et al. [88] assembled 548 CT scans from 4 datasets; Zhang et al. [88] assembled 1,155 CT scans from 7 datasets.
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 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 $D = \{(X_1, Y_1), (X_2, Y_2), ..., (X_N, Y_N)\}$, there are a total of $K$ unique classes. For $\forall n \in [1, N]$, if the presence of $\forall k \in [1, K]$ classes in $X_i$ 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 $D = \{D_1, D_2, ..., D_M\}$, 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 $F_{\theta}(X_n) = P_n$, $n \in [1, N], k \in [1, K]$, where the prediction $P_n$ is one-hot encoding with length $k$.

- **Solution #2** [88, 32, 99] aims to solve $F_{\theta}(X_n, u_k) = P_n, n \in [1, N], k \in [1, K]$, 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 in-depth in the following sections.

### 3.2. CLIP-Driven Universal Model

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

**Text branch.** Let \( w_k \) 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 \( \theta_k = MLP(w_k \oplus f) \), 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\(^4\). 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 \( P_k = \text{Sigmoid} \left( \left( F \ast \theta_{k1} \right) \ast \theta_{k2} \ast \theta_{k3} \right) \), where \( \theta_{k1}, \theta_{k2}, \theta_{k3} \) 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 \( 4 \times 10^{-4} \), momentum of 0.9 and decay of \( 1 \times 10^{-5} \) on multi-GPU (4) with DDP. The framework is implemented in MONAI 0.9.0\(^5\). The five-fold cross validation strategy is performed. We select the best model in each fold by evaluating the validation best metrics. Models are trained on eight NVIDIA RTX A5000 cards.

### 4.1. Organ Segmentation on MSD and BTCV

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

---

3Our framework design is conceptually similar to Segment Anything Model (SAM) [33], which is a concurrent study of ours in computer vision. By leveraging CLIP embedding as a prompt within our Universal Model, we are able to generate highly accurate masks for organs and tumors of interest, as opposed to producing masks for arbitrary objects.

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

5https://monai.io/

6lecathon-10.grand-challenge.org/evaluation/challenge/leaderboard/
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 9). 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, pancreas, and tumors. The quality of 19 other organs and 6 tumors has not been compared to human annotations because there is no publicly available CT scans have been annotated by multiple independent groups on these objects.
Table 3. 5-fold cross-validation results on BTCV. For a fair comparison, we did not use model ensemble during the evaluation. All experiments are under the same data splits, computing resources, and testing conditions. Universal Model achieves the overall best performance, yielding at least +3.9% DSC improvement over the state-of-the-art method.

<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>
<th>Avg.</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>88.96</td>
<td>82.49</td>
<td>73.54</td>
<td>75.48</td>
<td>66.09</td>
<td>80.76</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>
<td>80.94</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>
<td>81.22</td>
</tr>
<tr>
<td>UNEtR [22]</td>
<td>94.91</td>
<td>92.10</td>
<td>93.12</td>
<td>76.98</td>
<td>74.01</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>
<td>81.43</td>
</tr>
<tr>
<td>nuU-Net [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>
<td>82.01</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><strong>64.98</strong></td>
<td>82.06</td>
<td></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.

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 more 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 × 96 × 96. The size of each circle indicates the number of parameters (#P). In the inference, Universal Model is faster than nnU-Net (2nd best in performance) and Swin UNETR (3rd best) by 19× and 6× measured by FLOPs, respectively.

Figure 8. Expansibility: flexible backbones. Universal Model can be expanded to CNN-based (e.g., U-Net [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 anatomical relationship and to learn a structured feature embedding. Furthermore, we conduct ablation study with various embedding to replace the CLIP embedding including BioLinkBERT embedding [8], and the results are shown in Appendix Table 14. We can see that the CLIP-based embedding can significantly improve the performance comparing with conventional one-hot labels (DoDNet [88]) and text-only pre-trained embedding (BioLinkBERT [83]).
stream tasks in the TotalSegmentator dataset (see Table 6). The highest DSC compared with other pre-training methods is achieved with 86.49%, 89.57%, 94.43% and 88.95% for four downstream datasets. Through pre-training by assembly dataset directly and fine-tuning results, Universal Model can substantially outperform the previous methods on 3D-IRCADb and JHH with a DSC improvement of 5% and 4%, respectively (see Table 5). 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).

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. Compared with dataset-specific models, Universal Model has trained on the order of magnitude more diverse CT scans, therefore demonstrating significantly better generalizability. We conduct the evaluation on 3D-IRCADb (public) and JHH (private), which are absolutely not seen in the training. Universal Model substantially outperforms the previous methods on 3D-IRCADb and JHH with a DSC improvement of 5% and 4%, respectively (see Table 5).

Transferability: fine-tuning results. Universal Model can serve as a powerful pre-training model for segmentation. Through pre-training by assembly dataset directly and fine-tuning to other datasets, the Universal Model achieves the highest DSC compared with other pre-training methods with 86.49%, 89.57%, 94.43% and 88.95% for four downstream tasks in the TotalSegmentator dataset (see Table 5).

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 Wenzuan Li for their constructive suggestions at several stages of the project.
References


[20] Nicholas Heller, Niranj Sathianathan, Arveen Kalapara, Edward Walczak, Keenan Moore, Heather Kaluzniak, Joel


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


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


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


