CancerUniT: Towards a Single Unified Model for Effective Detection, Segmentation, and Diagnosis of Eight Major Cancers Using a Large Collection of CT Scans

Jieneng Chen[1,2*] Yingda Xia[1*] Jiawen Yao[1,3] Ke Yan[1,3] Jianpeng Zhang[1,3] Le Lu[1] Fakai Wang[1]
Bo Zhou[1] Mingyan Qiu[1,3] Qihang Yu[2] Mingze Yuan[1,3] Wei Fang[1,3] Yuxing Tang [1] Minfeng Xu[1,3]
Jian Zhou[4] Yuqian Zhao[5] Qifeng Wang[5] Xianghua Ye[6] Xiaoli Yin[7] Yu Shi[7] Xin Chen[8,9]
Jingren Zhou[1,3] Alan Yuille[2] Zaiyi Liu[8,9*] Ling Zhang[1]
[1]DAMO Academy, Alibaba Group [2]Johns Hopkins University
[3] Hupan Lab, 310023, Hangzhou, China [4]Sun Yat-sen University Cancer Center
[6]The First Affiliated Hospital of Zhejiang University [7] Shengjing Hospital of China Medical University
[5]Sichuan Cancer Hospital [8]Guangdong Provincial People's Hospital
[9] Guangdong Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application

Abstract

Human readers or radiologists routinely perform full-body multi-organ multi-disease detection and diagnosis in clinical practice, while most medical AI systems are built to focus on single organs with a narrow list of a few diseases. This might severely limit AI's clinical adoption. A certain number of AI models need to be assembled non-trivially to match the diagnostic process of a human reading a CT scan. In this paper, we construct a Unified Tumor Transformer (CancerUniT) model to jointly detect tumor existence & location and diagnose tumor characteristics for eight major cancers in CT scans. CancerUniT is a query-based Mask Transformer model with the output of multi-tumor prediction. We decouple the object queries into organ queries, tumor detection queries and tumor diagnosis queries, and further establish hierarchical relationships among the three groups. This clinically-inspired architecture effectively assists inter- and intra-organ representation learning of tumors and facilitates the resolution of these complex, anatomically related multi-organ cancer image reading tasks. CancerUniT is trained end-to-end using a curated large-scale CT images of 10,042 patients including eight major types of cancers and occurring non-cancer tumors (all are pathology-confirmed with 3D tumor masks annotated by radiologists). On the test set of 631 patients, CancerUniT has demonstrated strong performance under a set of clinically relevant evaluation metrics, substantially outperforming both multi-disease methods and an assembly of eight single-organ expert models in tumor detection, segmentation, and diagnosis. This moves one step closer towards a universal high performance cancer screening tool.

*Correspondence to Jieneng Chen (jienengchen01@gmail.com), Yingda Xia (yingda.xia@alibaba-inc.com), and Zaiyi Liu (zyliu@163.com)

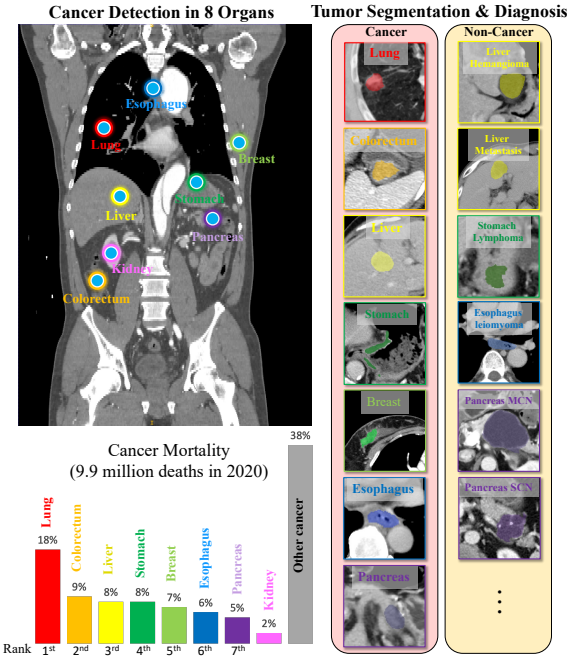

Figure 1. We aim at cancer and non-cancer detection, segmentation, and diagnosis in eight major organs via CT scan. Seven of our eight targeted cancers rank the top seven in terms of mortality.

1. Introduction

Cancer, a leading cause of death in the world, continues to thwart human life expectancy and cause huge soci-

etal burdens despite significant progress in medical research [52, 40]. Medical imaging is a powerful tool for detection and diagnostic examination of cancer and is widely used in clinical practice everywhere. The daily work of radiologists in reading and interpreting cancer imaging findings includes three main clinical tasks: detection, quantification, and diagnosis [3]. Since Computed Tomography (CT) body scans are very common (nearly 80% in all CT exams) [37] and each CT scan can have hundreds of image slices, the miss-detection and miss-diagnosis of cancer are the pain points in the radiology workflow. Human readers statistically tend to have high specificity but low sensitivity for tagging and reporting various anomalies or diseases.

Computer-aided detection (CADe) and diagnosis (CADx) can assist radiologists and oncologists to improve the tumor detection rate and diagnosis accuracy [14, 3]. With the development of deep learning and Convolutional Neural Networks (CNNs), CAD algorithms have met or exceeded expert-level performance in some specific applications [32, 27, 12, 2]. However, most CAD expert systems focus on dealing with single organ diseases [45], e.g., pancreas [76, 71], liver [23, 11], lung [2, 25], or kidney tumors [19], while radiologists in turn, must be responsible for all possible diseases and radiologically significant findings [34]. For example, for an abdominal CT examination that is initially targeted for the gallbladder, even when there are no clear prior indications (e.g., abdominal pain), all visible organs in the entire abdomen from CT imaging need to be carefully inspected by a radiologist. Therefore, the role of current CAD tools is still (very) limited in clinical practice, far from functioning as universally as a human reader. A versatile CAD tool that can perform (many) more critical medical tasks would be more clinically desirable and thus is in high demand [39].

Despite notable progress in multi-organ segmentation, it should be noted that detecting and diagnosing multiple cancers is considerably more difficult than segmenting organs alone due to several factors: (1) tumors have a variety of types, appearances and size, making it hard to be detected. (2) tumor detection requires differentiation of tumors from normal tissue within an organ, which is more challenging than the differentiation of organs from the background in organ segmentation. (3) the diagnosis of cancer involves the fine-grained categorization of tumors, which necessitates a high level of expertise and specialized training. Aiming at solving the universal lesion detection problem in CT scans, DeepLesion is a recent pioneering publicly available dataset [62, 61, 59], and despite much follow-up work, most cancer detection, quantification, and diagnosis solutions derived from DeepLesion dataset [62] are still insufficient in the following aspects. First, the data size and patient number of a single disease can be small, and some major cancer types (e.g., esophagus, stomach, and colorectum) are scarce, resulting in relatively high false positive and sub-optimal detection rates. Second, voxel-level tumor annotation, perhaps as 3D masks (requiring a high level of clinical expertise and are very tedious to label) are not available, making the necessary precise 3D quantification difficult (if not impossible). Third, the pathological gold standard of confirming tumor types is unavailable and, therefore, impossible to distinguish between malignant and benign lesions. Recent clinical validations of two multi-disease detection AI systems [36, 57] found that ruling out irrelevant CAD findings (i.e., false positives and lesions without adequate malignancy assessment) was very time-consuming and confusing for radiologists. These observations clearly indicates the essential limitations of applying DeepLesion dataset [62] to positive clinical impacts.

In this paper, we curate a large (abdominal and chest) CT image dataset including eight major cancers (from the top seven cancers with the highest mortality in the world [40]: lung, colorectum, liver, stomach, breast, esophagus, and pancreas, plus a public kidney dataset [19]) of total 10,673 patients (Of these, breast cancer has the fewest, with 478 patients; lung cancer has the most, with 2,402 patients. In addition, there are 1,055 normal controls). All tumor types (and subtypes) of the seven organs are confirmed by either surgical or biopsy pathology and recorded as gold standard labels, where full spectrum of all tumor subtypes are offered for four organs. All confirmed tumors in CT scans are manually segmented or delineated in 3D by board-certified radiologists who are specialized in the particular organ or disease types. To our knowledge, previous datasets with similar tumor characteristics only cover a single disease at the scale of hundreds of patients, such as the pancreatic tumor [71] and kidney tumor datasets [19]. The curation of our new 8-cancer dataset is a major step towards building a universal multi-cancer imaging reading AI model – with the hope to reach a performance level comparable with radiologists specializing in different cancer types – for assisting radiologists and general clinicians in precision detection, quantification, and diagnosis. Fig. 1 shows our goal for cancer and non-cancer detection, segmentation, and diagnosis in eight major organs via CT scans.

On the other hand, we propose a new clinically interpretable computing architecture, named Unified Tumor Transformer (CancerUniT). In general, CancerUniT is a single unified model that simultaneously solves the tasks of multi-tumor detection, segmentation and diagnosis in a semantic segmentation manner. Our motivations are: (1) the organs, cancers and non-cancer tumors are interrelated in both appearance similarity and human anatomical constraints, e.g., HCC (a major malignant liver cancer) and cyst (benign lesion) both occur inside the liver with textual and other visual differences, while HCC and PDAC (a type of pancreatic cancer) should appear in two different organs

but their clinical characteristics are both malignant carcinoma; (2) a unified learning of multi-organs-tumors could reduce the performance uncertainty and architectural complexity in assembling multiple single models, e.g., different predictions of the same intended object or finding by multiple models. To collaboratively model such differences and connections or dependencies, we propose a novel representation learning method that represents each organ and tumor as an object query of the Transformer in a semantic hierarchy. The object queries are divided into organ queries, tumor detection queries and diagnosis queries, and we establish a query hierarchy based on the clinical meaning of the queries. This design will explicitly encourage the queries to learn the inter-organ and intra-organ relationships to solve the clinically sophisticated multi-cancer tumor recognition tasks.

CancerUniT is trained and tested on our curated dataset. CancerUniT outperforms the DeepLesion model, the ensemble of single-organ expert models and unified baseline models (trained on our data). Compared to the DeepLesion model, CancerUniT has a 29.3% higher sensitivity and a large margin of 77.5% higher specificity in tumor detection. Compared to an ensemble of individually trained single-organ nnUNet models, CancerUniT has an average improvement of 6.7% in tumor detection sensitivity, 2.8% in diagnostic accuracy, and 3.9% in Dice segmentation score across all the organs; On normal patients, CancerUniT has an improvement of 22.5% in specificity (ours 81.7% vs. nnUNet 59.2%); CancerUniT is 4.5 times faster in testing speed. In comparison to a unified nnUNet model dealing with all eight organs, CancerUniT leads by 5.3% in lesion detection sensitivity, 6.7% in diagnostic accuracy, 2.8% in specificity, and 2.7% in Dice segmentation score. The improvements indicate that the different type of tumors have mutual correlations and the design of CancerUniT successfully captures this clinical relationship for enhanced tumor representation learning. The high performance of CancerUniT also sheds light on its clinical potential for real-world multi-cancer detection, segmentation, and diagnosis.

2. Related Work

CADe and CADx. CADe normally refers to the computer-aided localization process of lesions in 2D/3D medical images and CADx subsequently diagnoses lesions or findings as either malignant or benign [3] and assigning more potential tumor characteristics. Along with advances in deep learning, quantitative CADe performances matching or beyond medical domain experts are reported in several specialized single-organ clinical applications: breast cancer screening [32], lung cancer detection [2], retinal disease referral [12], skin disease diagnosis [27] and so on.

Tumor detection, segmentation, and diagnosis in CT via CNNs. CNNs have been widely applied to detect, segment, and diagnose cancers/tumors in CT scans. Lung nodule detection in low-dose CT [43] is the recommended lung cancer screening protocol where some promising results are discussed [2, 21, 55, 72]. Image segmentation networks [29, 8, 35, 24, 74] are well-adopted under the per-pixel classification setting and a segmentation model is designed to predict the probability distribution over all possible categories or labels per pixel (as a structured dense prediction problem). Segmenting abdominal organs and detecting tumor by segmentation principles [51, 1, 4, 19, 70, 67, 20], serve a key role towards fully-automated tumor detection [76, 53, 66, 54], differential diagnosis and reporting [71]. Despite their promising performance, these approaches are often specialized to focus only on a single organ. Multi-organ segmentation [30, 26, 50, 69, 13, 73] are emerging, but the degree of difficulty involved in multi-cancer detection and diagnosis is considerably greater than that of organ-level. DeepLesion [62] attempts to tackle the universal lesion detection task in CT scans, but their derived lesion detection methods [61, 59, 58] and several follow-up work [63, 31, 60, 42, 41] so far have reported mostly moderate multi-class lesion detection performance. Distinguishing between malignant and benign lesions in multi-class tumor setting is still far from a clinical reality.

Transformers [46] have advanced the state-of-the-art performance in various computer vision tasks [16, 5, 28, 7, 17, 75, 44], by capturing global interactions between image patches and having no built-in inductive prior. The success of Transformer has also been witnessed in medical image detection and segmentation [6, 56, 18]. With the recent progress in transformers [5, 48], a new variant called mask Transformers has been proposed, where segmentation predictions are represented by a set of query embeddings with their own semantic class labels, generated through the conversion of query embedding to mask embedding vectors followed by multiplying with the image features. The essential component of mask transformers is the decoder which takes object queries as input and gradually transfers them into mask embedding vectors. Recent works [38, 48, 10, 9] inspire us to represent tumor in the medical domain as the class query [38] within the Transformer formulation. In this paper, we propose a novel semantic hierarchical representation to exploit the relationship in detection, diagnosis and differentiation among eight main tumors and their sub-types from a large dataset of CT scans collected from both healthy subjects and patients with cancers.

3. Method

In this section, we first define the problem of tumor detection, segmentation, and diagnosis from an image semantic segmentation perspective in Sec. 3.1. We then describe the overview of query-based mask Transformer and how we integrate it as our segmentation decoder in Sec. 3.2. After

that, we introduce the proposed Unified tumor Transformer (CancerUniT) in Sec. 3.3, which represents tumors by a semantic query hierarchy, solving the tumor detection, segmentation, and diagnosis in a unified manner.

3.1. Problem Definition

We focus on three tasks in images, i.e., tumor detection, segmentation, and diagnosis. Tumor detection aims to locate the presence of target types of tumors. Tumor segmentation aims to provide per-pixel annotation of the tumor region. Tumor diagnosis aims to classify the specific tumor subtype of a detected tumor. We denote $\mathbf{s}_o$ as a set of organs, and $\mathbf{s}_t$ as a set of tumors. Specifically in our dataset,

$\mathbf{s}_o$	{breast, lung, kidney, pancreas, esophagus, liver, stomach, colorectum}
$\mathbf{s}_t$	{breast cancer, lung cancer, colorectal cancer, pancreas PDAC, pancreas nonPDAC, liver HCC, liver ICC, liver metastatis, liver hemangioma, stomach GC, stomach nonGC, esophagus EC, esophagus nonEC, kidney tumor/cyst}

We propose to solve these three tasks with a semantic segmentation framework, in which we assign each voxel in the CT scan with a semantic label $k \in \mathbf{s}_o \cup \mathbf{s}_t$ and the total number of classes is $K = |\mathbf{s}_o| + |\mathbf{s}_t|$. The tumor detection, segmentation, and diagnosis are then evaluated based on the semantic segmentation results.

3.2. Basis: Query-based Mask Transformers

Although Transformers have been used for medical image segmentation as feature extractors, the query-based mask Transformer [38, 48, 10] decoder is rarely explored in medical images. Query-based mask Transformer aims to decode the pixel-level features (usually from a CNN backbone) with object queries. Our method is based on this design and here we provide an overview of its basic components. **Query initialization.** A set of K learnable class queries (*i.e.*, embeddings) $\mathbf{q} = [\mathrm{q}_1, ..., \mathrm{q}_K] \in \mathbb{R}^{K \times d}$ is defined where K is the number of classes and d is the query dimension. Each class query is initialized randomly and assigned to a single semantic class.

Query interaction via Transformer. The queries are updated through multi-head cross-attention, multi-head self-attention, and feedforward network [46]. The multi-head cross-attention between queries and image features is computed to update queries conditioned on image features. The multi-head self-attention allows queries to interact with each other.

Decode queries to segmentation. The class query $\mathbf{q}$ is processed jointly with 3D image features $\mathbf{F} \in \mathbb{R}^{d \times D \times H \times W}$ by the decoder. K masks can be generated by computing the scalar product between L2-normalized image features $\mathbf{F} \in \mathbb{R}^{d \times D \times H \times W}$ and class queries $\mathbf{q} \in \mathbb{R}^{K \times d}$. The set of class masks is computed as:

$$\mathbf{M} = \mathbf{q} \times \mathbf{F} \tag{1}$$

where $\mathbf{M} \in \mathbb{R}^{K \times D \times H \times W}$ is K mask predictions and will be followed by a softmax to obtain the final pixel-wise class probability map in the task of semantic segmentation.

3.3. CancerUniT: Unified Tumor Transformer

We introduce the novel unified tumor transformer (See Fig. 2), which includes semantic query hierarchy for tumor representation, UNet backbone for feature extraction, Transformer for query interaction, and dual-task query decoding for tumor detection task and cancer diagnosis task.

3.3.1 Query Hierarchy

We propose a novel tumor representation via a semantic hierarchy of queries, including shared, detection, and diagnosis queries. By this design, tumors are represented as queries and a "detection-to-diagnosis hierarchy" is established based on the semantic relationship of tumors.

We hereby divide the segmentation targets $\mathbf{s}_o \cup \mathbf{s}_t$ into three non-overlapping groups, i.e., $\mathbf{m}$, $\mathbf{n}$, and $\mathbf{s}$. $\mathbf{m}$ consists of m general tumor categories that requires further diagnosis. The i^{th} element $\mathbf{m}_i$ can be further categorized into the sub-classes of $\mathbf{n}_i$ with a number of n_i sub-classes. $\mathbf{s}$ consists the rest of the targets including eight organs and the four cancers that do not require diagnosis in our data.

Shared query. We define a set of shared query $\mathbf{s} \in \mathbb{R}^{s \times d}$ to represent the shared classes in both detection task and diagnosis task. The shared classes include the 8 organ classes, and 4 tumor classes without other sub-types.

$\mathbf{s}$	$\mathbf{s}_o \cup$ {lung cancer, breast cancer, colorectal cancer, kidney tumor/cyst}
$\mathbf{m}$	{pancreas tumor, liver tumor, stomach tumor, esophagus tumor}
$\mathbf{n}$	{{PDAC, nonPDAC}, {HCC, ICC, metastasis, hemangioma}, {GC, nonGC}, {EC, nonEC)}}

Detection query. We denote $\mathbf{A} \in \mathbb{R}^{m \times d}$ as detection queries with m specifying the number of queries. Each detection query corresponds to the general tumor class of an organ, which requires further diagnosis.

Diagnosis query. The cancer diagnosis relies on fine-grained tumor categorization. Similarly, we denote a feature embedding $\mathbf{B} \in \mathbb{R}^{n_i \times d}$ as diagnosis queries with n_i specifying the number of diagnosis classes for tumor $\mathbf{m_i}$. A group of n_i diagnosis queries refers to $|\mathbf{n_i}|$ tumor sub-types $\mathbf{n_i}$ occurred in organ i, and totally we have $n = \sum_{i=1}^{|\mathbf{n}|} n_i$ diagnosis queries in this work.

Detection-to-Diagnosis hierarchy via linear projection. Inspired by the clinical practice of detection-then-diagnosis and given the fact that diagnosis queries are subtypes of detection queries, we aim to build a graph treating the detection queries as parent nodes and diagnosis queries

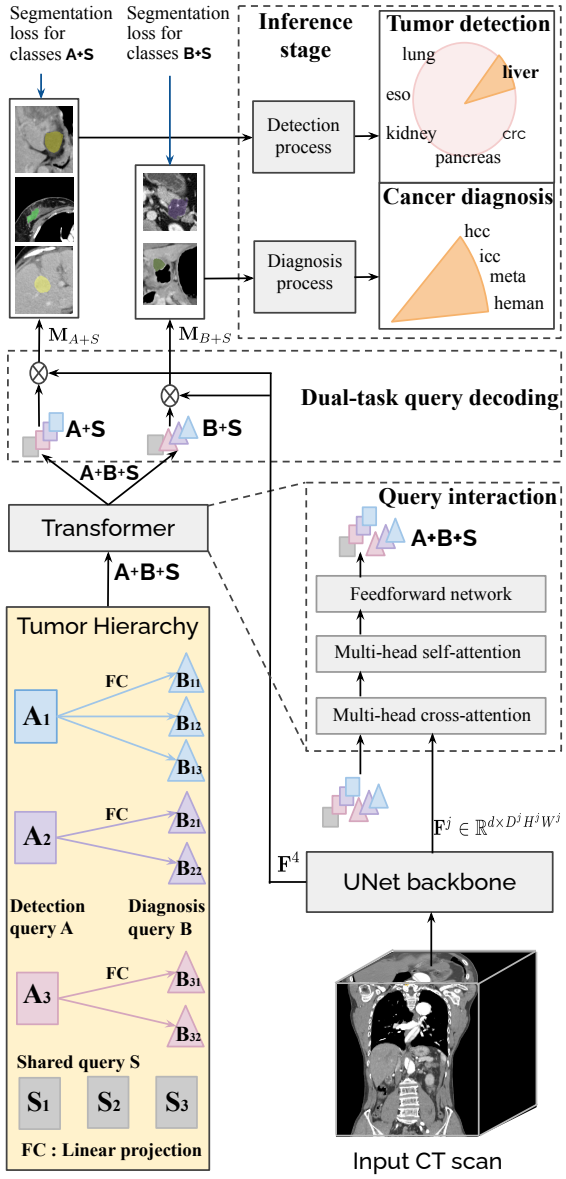

Figure 2. Overview of Unified Tumor Transformer (CancerUniT). We first represent tumor as queries $\mathbf{A}$ and $\mathbf{B}$ (*i.e.* feature embedding), and then build the query hierarchy from $\mathbf{A}$ to $\mathbf{B}$ via linear projection $\mathbf{FC}$ according to the tumor sub-type relationship. The tumor queries interact and are updated in a Transformer decoder with input of UNet features $\mathbf{F}$. Dual-task query decoding is performed to generate semantic segmentation map for two tasks. The detection task focuses on major tumor classes while diagnosis task is supervised by fine-grained cancer sub-types. In inference stage, the dual-task tumor segmentation maps are post-processed separately to produce multi-classes tumor instances for tumor detection and cancer diagnosis.

as children nodes. In this way, the model is able to learn the hierarchical representation of tumors explicitly.

To build the semantic hierarchical relationship, we project a detection query $\mathbf{A}_i \in \mathbb{R}^{1\times d}$ into diagnosis queries $\mathbf{B}_i \in \mathbb{R}^{1\times n_i d}$ via a linear projection layer with matrix $\mathbf{W}_i \in \mathbb{R}^{n_i d\times d}$. The detection-to-diagnosis procedure is formulated as:

$$\mathbf{B}_i = \mathbf{A}_i \times \mathbf{W}_i^T \tag{2}$$

where $\mathbf{B}_i = [\mathbf{B}_{i,1}, \mathbf{B}_{i,2}, ...\mathbf{B}_{i,n_i}]$ and the subgroup capacity $n_i = |\mathbf{n_i}|$ represents the number of subtypes for tumor $\mathbf{m_i}$. The detection queries $\mathbf{A}$ are learnable parameters that are random initialized, while diagnosis queries $\mathbf{B}$ are feature embedding conditioned on detection queries.

3.3.2 Meta Architecture

The proposed architecture includes a UNet backbone for feature extraction, a Transformer for query interaction, and a dual-task query decoding stage to generate segmentation masks. Detailed model instantiation is in Appendix.

nnUNet backbone for feature extraction. We adopt nnUNet [24] as the backbone to extract multi-scale features $\mathbf{F} = [\mathbf{F}^1, \mathbf{F}^2, \mathbf{F}^3, \mathbf{F}^4]$ where $\mathbf{F}^j \in \mathbb{R}^{d\times D^j H^j W^j}$ is the j-th layer feature map after projecting to number of channel d and flattening the spatial dimension D^j, H^j, and W^j.

Transformer for query interaction. We use the standard Transformer decoder [46] with input of UNet features $\mathbf{F}^j$ and queries $[\mathbf{A}^j, \mathbf{B}^j, \mathbf{S}^j]$ at the j-th layer. The Transformer is stacked by three Transformer layers, each of which contains a multi-head cross-attention, a multi-head self-attention, and a feed-forward network. The concatenated query $[\mathbf{A}, \mathbf{B}, \mathbf{S}]$ is updated via the cross attention (denoted as CA) between the queries and the image feature $\mathbf{F}^j$, as well as the query self-attention (denoted as SA). The query interaction is written as:

$$\mathbf{A}^j, \mathbf{B}^j, \mathbf{S}^j = SA(CA([\mathbf{A}^{j-1}, \mathbf{B}^{j-1}, \mathbf{S}^{j-1}], \mathbf{F}^j)) \tag{3}$$

Dual-task query decoding. As there exists inclusiveness for the classes in $\mathbf{A}$ and $\mathbf{B}$, it is unlikely to decode them jointly if we would like to enforce multi-class exclusivity constraint (*e.g.* softmax). To better capture the class exclusivity, we propose the dual-task query decoding procedure that decodes queries $[\mathbf{A}, \mathbf{S}]$ and queries $[\mathbf{B}, \mathbf{S}]$ separately to perform dual-task semantic segmentation. The query decoding follows Eq. 1 with a softmax activation, written as:

$$\begin{aligned} \mathbf{M}_{A+S} &= softmax([\mathbf{A}, \mathbf{S}] \times \mathbf{F}^4) \\ \mathbf{M}_{B+S} &= softmax([\mathbf{B}, \mathbf{S}] \times \mathbf{F}^4) \end{aligned} \tag{4}$$

where $\mathbf{M}_{A+S}$ and $\mathbf{M}_{B+S}$ are the decoded voxel-wise semantic map for the detection task and the diagnosis task, respectively.

End-to-end training. Our method performs both major tumor segmentation and tumor subtype segmentation directly from CT scans, while vanilla methods only output subtype segmentation maps that are further merged to major tumor segmentation maps. In our work, the loss function

is the combination of cross-entropy loss and Dice loss [33], which are applied to both detection output $\mathbf{M}_{A+S}$ and diagnosis output $\mathbf{M}_{B+S}$ to enforce the similarity with their corresponding targets.

Inference. End-to-end inference of dual-task segmentation is enabled simultaneously. For the detection process, the tumor segmentation map from $\mathbf{M}_{A+S}$ is extracted to generate tumor instances (*i.e*. connected components) with tumor class label i. If the predicted tumor instance has overlaps with the ground-truth tumor, the patient is detected with tumor class i. For the diagnosis process, we do similar tumor instance extraction from $\mathbf{M}_{B+S}$, but each tumor instance is identified as one specific tumor subtype. The patient-level cancer diagnosis category is decided by the tumor subtype with the largest connected component.

4. Experiments

4.1. Experiment Setup

Dataset description. Our 8-cancer CT dataset, which includes seven in-house tumor datasets (collected from five hospitals), one publicly available kidney tumor dataset [19], and a normal control dataset, is composed of 10,673 contrast-enhanced CT volumes (all in venous phase, except for lung and CT angiography being arterial phase), each from one unique patient. These CT volumes are acquired before treatment. All cancers (and tumor subtypes) in the seven in-house datasets are confirmed by pathology, with four datasets having a full spectrum of tumor subtypes, i.e., liver (4 subtypes), stomach (6), esophagus (4), and pancreas (9). The normal controls consist of 934 abdominal CT and 121 CT angiography (CTA) scans. Some of the datasets for single organs have been involved in our previous publications for other precision oncology research purposes [71, 65, 66, 67, 22, 49, 15, 64].

Tumors in each organ dataset are manually segmented slice-by-slice on CT images by radiologists who provide the data and specialize in the specific disease using either ITK-SNAP [68] or our in-house developed CT annotation tool – CTLabler [47]. During annotation, the radiologists also refer to the other CT phases (e.g., arterial, delay), contrast-enhanced MRI, and radiology/surgery/pathology reports if necessary. All organs are segmented/delineated automatically: the breast is by a nnUNet model trained on 213 additional breast cancer CT volumes with CTV (clinical target volume) masks; the other seven organs are by another nnUNet model trained on the Totalsegmentator dataset [51]. We randomly select 50 CT volumes from each tumor subtype (except for the liver tumor subtypes being 15 each, considering the relatively small liver data size), 50 abdominal, and 21 CTA volumes to form the test set. The remaining 10,042 CT volumes are used as the training set (Table 1).

Implementation details. All images were resampled to a spacing of $3 \times 0.8 \times 0.8$mm $(Z \times X \times Y)$. In the training stage, we randomly cropped sub-volumes of size of $48 \times 192 \times 192$ voxels from CT scans as the input. We employed the online data augmentation of nnUNet [24], including random rotation, scaling, flipping, Gaussian blurring, adding white Gaussian noise, adjusting brightness and contrast, simulation of low resolution, and Gamma transformation, to diversify the training set. The balanced sampling strategy was adopted to encourage model to sample different datasets and also different organ regions evenly. The batch size was set to 8, with 1 batch size per GPU on an 8-GPU machine. We adopted the AdamW optimizer and an initial learning rate of 3e-4. The baseline models were trained from scratch with 700 epochs, and the number of iterations per epoch equaled to training dataset size divided by the batch size. It took 40 GPU days to train a nnUNet from scratch on our dataset with Nvidia V100 GPUs. Due to huge cost, CancerUniT was trained based on the pre-trained nnUNet with a learning rate multiplier 0.1, and we trained 50 epochs. For fair comparison, we also kept tuning nnUNet for another 50 epochs besides 700 epochs whereas no performance improvement was observed.

In the inference stage, we employed the sliding window strategy, where the window size equals to the training patch size. In addition, Gaussian importance weighting and test time augmentation by flipping along all axes were also utilized to improve the robustness of segmentation.

Evaluation metrics. We consider the evaluation metrics from three aspects, including patient-level, lesion-level, and dense-level metrics. For patient-level evaluation, sensitivity and specificity are computed. Lesion-level precision and recall (do not consider the tumor type) are computed based on connected component analysis of tumor predictions. Tumor segmentation accuracy is assessed by the Dice coefficient.

Baselines. We compare our method to five baselines. (i) 8-nnUNet ensemble: An ensemble of 8 separately trained nnUNet [24] models. To solve overlapping tumor predictions, we extract the tumor connected components from 8 model predictions and merge them with the priority of tumor size. (ii) nnUNet: A unified nnUNet trained on our dataset as a multi-organ multi-tumor segmentation task. (iii) TransUNet: A leading Transformer model TransUNet [6] on medical image segmentation implemented in the nnUNet framework [24] and with the same settings as (ii). (iv) DeepLesion model: A universal lesion detection model [58] trained on the DeepLesion dataset [62]. (v) LENS (train on our data): A leading medical lesion detection algorithm LENS [58] trained on our dataset.

For a fair comparison, all segmentation-based models adopt fair data augmentations following nnUNet [24] and the same training techniques, while LENS and DeepLesion as detection-based methods adopt augmentations and training techniques following LENS [58]. All the models are

Table 1. Dataset description. A-F denote six different hospitals.

	Cancers				Full spectrum tumors										Normal controls		Total
	Breast	CRC	Kidney	Lung	Pancreas		Esophagus		Stomach		Liver				Abdomen	CTA	
Hospitals	A	A	public	A,B	C		B,D,E		A		C				A	F	
Subtypes	BC	CRC	KT	LC	PDAC	nonPDAC	EC	nonEC	GC	nonGC	HCC	ICC	Meta	Heman	-	-	
Train	428	746	249	2352	1315	727	1185	105	1117	273	284	31	99	147	884	100	10042
Test	50	50	50	50	50	50	50	50	50	50	15	15	15	15	50	21	631
Total	478	796	299	2402	1365	777	1235	155	1167	323	299	46	114	162	934	121	10673

Table 2. Patient-level tumor detection results. Sensitivity (%) and specificity (%) are reported. "DeepLesion model" is trained on DeepLesion dataset [61]) using the detection-based algorithm LENS [58]; "LENS (trained on our data)" is trained on our new 8-cancer dataset.

Model	Sensitivity (%)									Specificity (%)		
	Br.	Crc.	Kid.	Lung	Pan.	Eso.	St.	Liv.	Average	Abd.	CTA	Average
8-nnUNet ensemble	**96.0**	74.0	94.0	74.0	93.0	83.0	92.0	86.7	86.6	80.0	9.5	59.2
DeepLesion model [62]	78.0	38.0	86.0	76.0	82.0	34.0	30.0	88.3	64.0	6.0	0.0	4.2
LENS (train on our data) [58]	82.0	62.0	76.0	50.0	89.0	72.0	72.0	75.0	72.3	70.0	52.4	64.8
nnUNet [24]	90.0	82.0	92.0	94.0	94.0	76.0	91.0	85.0	88.0	92.0	47.6	78.9
TransUNet [6, 24]	94.0	86.0	94.0	94.0	94.0	81.0	**94.0**	88.3	90.7	94.0	47.6	80.3
Ours	94.0	**92.0**	**94.0**	**94.0**	**95.0**	**89.0**	93.0	**95.0**	**93.3**	**96.0**	**47.6**	**81.7**

Table 3. Class-agnostic lesion instance-level detection results. We treat eight types of tumors as one class. The numbers of FN, TP, FP lesions, precision and recall are reported. The total number of ground-truth lesions in the test set is 767. Note one patient might have several lesions in the 560 patients with tumors, and a ground-truth lesion might be matched with multiple TP components.

Model	FN	TP	FP	Precision	Recall
8-nnUNet ensemble	209	568	1060	34.9%	72.8%
DeepLesion model [62]	376	649	5345	10.8%	51.0%
LENS (train on our data) [58]	267	602	875	40.8%	65.2%
nnUNet [24]	223	557	585	48.8%	70.9%
TransUNet [6, 24]	169	648	726	47.2%	**77.9%**
Ours	192	592	508	**53.8%**	75.0%

trained to be converged.

4.2. Main Results

Patient-level tumor detection per organ. This task aims at the evaluation of whether the model can correctly localize and identify an existing tumor (agnostic of subtypes) or generate false positive tumor predictions in the normal controls. For example, if a patient has a tumor annotated in the liver, a true positive prediction means that the model predicts a liver tumor that overlaps (Dice > 0) the ground-truth tumor annotation. We report the sensitivity for each organ and the specificity for normal controls in the test set.

As shown in Table 2, our model outperforms all the baseline models in terms of average sensitivity and specificity. Compared to the 8-nnUNet ensemble, our model has substantial improvement in the sensitivity of detecting colorectum (+18%), lung (+20%), and liver (+8%) tumors, and the overall specificity (+21%). We also observe improvements in these organs of other unified models, i.e., nnUNet and TransUNet, which demonstrate that the unified training of multi-organ multi-tumor segmentation will benefit almost every separate task, except for breast tumor (-2%). Without seeing other organs and tumors, the separately trained models have many more false positives than unified models, with a much lower specificity of 59.2%.

Without seeing our data, the DeepLesion [62] model has a moderate average sensitivity (64.0%) and low specificity (4.2%), hardly applicable to the real clinical scenario under such a high false positive rate. After training on our data with a leading lesion detection algorithm LENS [58], the sensitivity for colorectum, esophagus, and stomach, as well as the specificity are substantially improved; nevertheless, these are still lower than the segmentation-based models. These comparisons demonstrate that solving the tumor detection task as semantic segmentation is superior to using object detection methods.

Class-agnostic lesion-level tumor detection. In lesion or tumor-level evaluation, we combine all lesions into one class and extract the lesion instances from the segmentation masks of ground-truth and predictions to compute the overall precision and recall. If a predicted lesion instance mask overlaps a ground-truth lesion, we count this prediction as true positive. As shown in Table 3, our approach has the highest precision and recall among all the methods. Both the DeepLesion models and the 8-nnUNet ensemble models have a large number of false positives, resulting in low precision. Similar to patient-level results, semantic segmentation algorithms generally do better than object detection methods. Our model outperforms the unified nnUNet model by approximately 5% in precision and 4% in recall.

Tumor segmentation. This task focuses on the tumor segmentation quality, where our model still ranks as the top in segmentation Dice score per organ, as shown in Table 4. Here, we still ignore the subtype of the tumor and treat the tumors in the same organ with the same label. We only compare our model with the segmentation baselines, not the detection models (DeepLesion and LENS). Similar to tumor detection, the second best is the TransUNet model, and the

Table 4. Voxel-level tumor semantic segmentation results (in Dice coefficient %).

Model	Breast	Colorectum	Kidney	Lung	Pancreas	Esophagus	Stomach	Liver	average
8-nnUNet ensemble	0.623	0.474	0.728	0.415	0.690	**0.661**	0.420	0.703	0.589
nnUNet [24]	0.661	0.515	0.736	**0.548**	0.695	0.597	0.418	0.676	0.601
TransUNet [6, 24]	0.700	0.530	0.738	0.540	0.700	0.621	**0.444**	0.691	0.620
Ours	**0.702**	**0.533**	**0.739**	0.515	**0.702**	0.652	0.435	**0.743**	**0.628**

Table 5. Patient-level cancer diagnosis. The sensitivity (%) for each tumor subtype is reported. We categorize tumor subtypes as two classes of cancer and non-cancer tumors for pancreas, esophagus, and stomach datasets; and consider four major subtypes for liver dataset.

Model	Pancreas			Esophagus			Stomach			Liver					Average
	PDAC	nonPDAC	avg	EC	nonEC	avg	GC	nonGC	avg	HCC	ICC	Meta	Heman	avg	
8-nnUNet ensemble	88.0	74.0	81.0	92.0	32.0	62.0	94.0	28.0	61.0	**80.0**	60.0	**46.7**	86.7	**68.3**	68.1
nnUNet [24]	92.0	76.0	84.0	94.0	12.0	53.0	96.0	18.0	57.0	69.0	69.0	33.3	80.0	62.8	64.2
TransUNet [6, 24]	**94.0**	78.0	86.0	94.0	22.0	58.0	**96.0**	18.0	57.0	60.0	80.0	40.0	80.0	65.0	66.5
Ours	90.0	**84.0**	**87.0**	**94.0**	**36.0**	**65.0**	82.0	**48.0**	**65.0**	60.0	**80.0**	40.0	**86.7**	66.7	**70.9**

Table 6. Ablation study on the representation of tumor queries. Average detection sensitivity (%) and specificity (%), and voxel-level tumor Dice scores are reported.

	Sensitivity	Specificity	Dice
Plain	89.5	76.1	0.605
Parallel	90.1	78.9	0.608
Hierarchy (Ours)	93.3	81.7	0.628

Table 7. Efficiency comparison. CancerUniT is 4.5x faster and 8x lighter than the assembly of single-tumor expert models (8-nnUNet).

Model	Speed	Params
8-nnUNet ensemble	187s	246.24M
DeepLesion model [62]	17s	70.94M
LENS (train on our data) [58]	17s	70.94M
nnUNet [24]	22s	30.78M
TransUNet [6, 24]	25s	38.53M
Ours	42s	30.87M

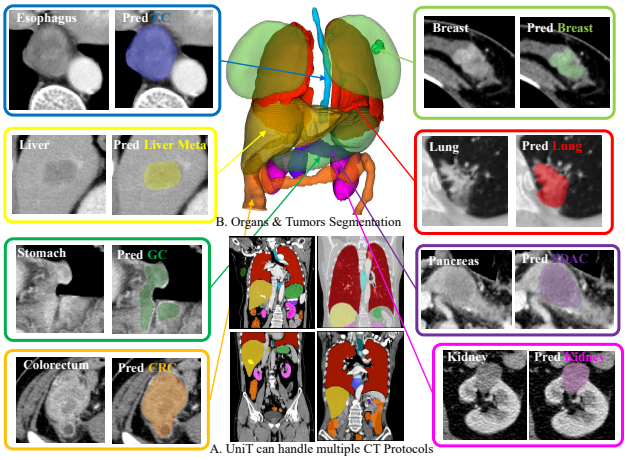

Figure 3. (A) Our model can handle multiple CT protocols representing the real-world clinical practice. (B) An example of the 3D masks of 8 organs and a breast tumor; and examples of 8 types of tumors being detected, segmented, and diagnosed by our CancerUniT.

unified nnUNet is better than its single counterpart. The improved performance of our model and TransUNet model illustrates that enhancing the CNN feature extraction with attentions will benefit multi-tumor segmentation. This observation is in line with our assumption that our query-based Transformer better explores the similarity between the inter-organ tumors, thus mutually improving the pixel-level texture differentiation of all tumors.

Tumor diagnosis. Our third evaluation focuses on the diagnostic ability to differentiate different types of tumors on the four organs, i.e., pancreas, esophagus, stomach, and liver, where we have tumor subtypes including cancer and non-cancer. As shown in Table 5, our method achieves the highest overall diagnosis performance of 70.9%. Different from previous tumor detection and segmentation results, the single expert model is the second best (68.1%), demonstrating its strong baseline performance on the diagnosis on single organs. The unified nnUNet model has a substantial performance drop (-4%) compared to its separately trained counterpart. We hypothesize that this is due to the difficulty of multi-task training. With only voxel-wise supervision, a vanilla unified nnUNet is hard to well recognize numerous subtypes of tumors for accurate diagnosis. In contrast, our model is capable of exploiting the relationship between different tumor diagnosis tasks with our query hierarchy, thus maintaining high performance and even improving over the single expert models.

Ablation study. We perform the ablation study on the representation of tumors (Table 6). We compare the other two representations: (1) parallel representation: the detection queries and diagnosis queries are organized as two groups in parallel, without structural connection. (2) plain representation: only diagnosis queries are used in our framework, while the prediction for a major tumor $\boldsymbol{m}_i$ in

the detection branch is directly obtained by merging several subtypes of tumors n_i in the diagnosis branch.

Efficiency. We compare the efficiency among various models in both inference speed and model size (number of parameters) as illustrated in Table 7. CancerUniT is 4.5x faster and 8x lighter than the assembly of single-tumor expert models.

Visual results in Fig. 3 shows that our model can handle multiple CT protocol, and is capable to detect, segment, and diagnose 8 types of major cancers. **Generalizability** to public dataset is shown in Supplementary.

5. Conclusion

In this paper, we propose a single unified tumor Transformer (CancerUniT) model to detect, segment and diagnose eight common cancers using 3D CT scans, for the first time. CancerUniT is a query-based Transformer and offers a novel clinically inspired hierarchical tumor representation, with a dual-task query decoding stage for segmentation mask generation. We curate a large collection of CT scans with high clinical quality from 10,673 patients, including eight major types of cancers and occurring non-cancer tumors (pathology-confirmed and manually annotated). Extensive quantitative evaluations have demonstrated the promising performance of our new model. This moves one step closer to a universal high performance cancer screening AI tool.

Acknowledgments. Jieneng Chen and Alan Yuille in this project were partially funded by a 2023 Patrick J. McGovern Foundation award.

References

[1] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022. 3

[2] Diego Ardila, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, 25(6):954–961, 2019. 2, 3

[3] Wenya Linda Bi, Ahmed Hosny, Matthew B Schabath, Maryellen L Giger, Nicolai J Birkbak, Alireza Mehrtash, Tavis Allison, Omar Arnaout, Christopher Abbosh, Ian F Dunn, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. *CA: A Cancer Journal for Clinicians*, 69(2):127–157, 2019. 2, 3

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

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 3

[6] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 3, 6, 7, 8

[7] Jie-Neng Chen, Shuyang Sun, Ju He, Philip HS Torr, Alan Yuille, and Song Bai. Transmix: Attend to mix for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12135–12144, 2022. 3

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 3

[9] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 3

[10] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 3, 4

[11] Chi-Tung Cheng, Jinzheng Cai, Wei Teng, Youjing Zheng, Yu-Ting Huang, Yu-Chao Wang, Chien-Wei Peng, Youbao Tang, Wei-Chen Lee, Ta-Sen Yeh, et al. A flexible three-dimensional heterophase computed tomography hepatocellular carcinoma detection algorithm for generalizable and practical screening. *Hepatology Communications*, 2022. 2

[12] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9):1342–1350, 2018. 2, 3

[13] Konstantin Dmitriev and Arie E Kaufman. Learning multi-class segmentations from single-class datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9501–9511, 2019. 3

[14] Kunio Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31(4-5):198–211, 2007. 2

[15] D Dong, M-J Fang, L Tang, X-H Shan, J-B Gao, F Giganti, R-P Wang, X Chen, X-X Wang, D Palumbo, et al. Deep learning radiomic nomogram can predict the number of lymph node metastasis in locally advanced gastric cancer: an international multicenter study. *Annals of Oncology*, 31(7):912–920, 2020. 6

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,

Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[17] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 3

[18] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2022. 3

[19] Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. *Medical Image Analysis*, 67:101821, 2021. 2, 3, 6

[20] Ahmed Hosny, Danielle S Bitterman, Christian V Guthier, Jack M Qian, Hannah Roberts, Subha Perni, Anurag Saraf, Luke C Peng, Itai Pashtan, Zezhong Ye, et al. Clinical validation of deep learning algorithms for radiotherapy targeting of non-small-cell lung cancer: an observational study. *The Lancet Digital Health*, 4(9):e657–e666, 2022. 3

[21] Xiaojie Huang, Junjie Shan, and Vivek Vaidya. Lung nodule detection in ct using 3d convolutional neural networks. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 379–383. IEEE, 2017. 3

[22] Yan-qi Huang, Chang-hong Liang, Lan He, Jie Tian, Cui-shan Liang, Xin Chen, Ze-lan Ma, and Zai-yi Liu. Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer. *Journal of Clinical Oncology*, 34(18):2157–2164, 2016. 6

[23] Yuankai Huo, Jinzheng Cai, Chi-Tung Cheng, Ashwin Raju, Ke Yan, Bennett A Landman, Jing Xiao, Le Lu, Chien-Hung Liao, and Adam P Harrison. Harvesting, detecting, and characterizing liver lesions from large-scale multi-phase CT data via deep dynamic texture learning. *arXiv preprint arXiv:2006.15691*, 2020. 2

[24] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. 3, 5, 6, 7, 8

[25] Roger Y Kim, Jason L Oke, Lyndsey C Pickup, Reginald F Munden, Travis L Dotson, Christina R Bellinger, Avi Cohen, Michael J Simoff, Pierre P Massion, Claire Filippini, et al. Artificial intelligence tool for assessment of indeterminate pulmonary nodules detected with CT. *Radiology*, page 212182, 2022. 2

[26] Ho Hin Lee, Yucheng Tang, Olivia Tang, Yuchen Xu, Yunqiang Chen, Dashan Gao, Shizhong Han, Riqiang Gao, Michael R Savona, Richard G Abramson, et al. Semi-supervised multi-organ segmentation through quality assurance supervision. In *Medical Imaging 2020: Image Processing*, volume 11313, pages 363–369. SPIE, 2020. 3

[27] Yuan Liu, Ayush Jain, Clara Eng, David H Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, et al. A deep learning system for differential diagnosis of skin diseases. *Nature Medicine*, 26(6):900–908, 2020. 2, 3

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 3

[29] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 3

[30] Xiangde Luo, Wenjun Liao, Jianghong Xiao, Jieneng Chen, Tao Song, Xiaofan Zhang, Kang Li, Dimitris N Metaxas, Guotai Wang, and Shaoting Zhang. Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *Medical Image Analysis*, 82:102642, 2022. 3

[31] Fei Lyu, Baoyao Yang, Andy J. Ma, and Pong C. Yueni. A segmentation-assisted model for universal lesion detection with partial labels. In *MICCAI*, 2021. 3

[32] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020. 2, 3

[33] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 6

[34] Perry J Pickhardt. Value-added opportunistic CT screening: State of the art. *Radiology*, 2022. 2

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[36] Johannes Rueckel, Jonathan I Sperl, Sophia Kaestle, Boj F Hoppe, Nicola Fink, Jan Rudolph, Vincent Schwarze, Thomas Geyer, Frederik F Strobl, Jens Ricke, et al. Reduction of missed thoracic findings in emergency whole-body computed tomography using artificial intelligence assistance. *Quant Imaging Med Surg*, 11:2486–98, 2021. 2

[37] Aaron Sodickson, Pieter F Baeyens, Katherine P Andriole, Luciano M Prevedello, Richard D Nawfel, Richard Hanson, and Ramin Khorasani. Recurrent CT, cumulative radiation exposure, and associated radiation-induced cancer risks from CT of adults. *Radiology*, 251(1):175, 2009. 2

[38] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021. 3, 4

[39] Ronald M Summers. Road maps for advancement of radiologic computer-aided detection in the 21st century. *Radiology*, 229(1):11–13, 2003. 2

[40] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021. 2

[41] Youbao Tang, Jinzheng Cai, Ke Yan, Lingyun Huang, Guotong Xie, Jing Xiao, Jingjing Lu, Gigin Lin, and Le Lu. Weakly-supervised universal lesion segmentation with regional level set loss. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 515–525. Springer, 2021. 3

[42] You-Bao Tang, Ke Yan, Yu-Xing Tang, Jiamin Liu, Jin Xiao, and Ronald M Summers. Uldor: a universal lesion detector for ct scans with pseudo masks and hard negative example mining. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 833–836. IEEE, 2019. 3

[43] National Lung Screening Trial Research Team et al. The national lung screening trial: overview and study design. *Radiology*, 258(1):243, 2011. 3

[44] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 3

[45] Kicky G van Leeuwen, Steven Schalekamp, Matthieu JCM Rutten, Bram van Ginneken, and Maarten de Rooij. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *European Radiology*, 31(6):3797–3804, 2021. 2

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 4, 5

[47] Fakai Wang, Chi-Tung Cheng, Chien-Wei Peng, Ke Yan, Min Wu, Le Lu, Chien-Hung Liao, and Ling Zhang. Multi-sensitivity segmentation with context-aware augmentation for liver tumor detection in CT. *in submission*. 6

[48] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5463–5474, 2021. 3, 4

[49] Shuo Wang, He Yu, Yuncui Gan, Zhangjie Wu, Encheng Li, Xiaohu Li, Jingxue Cao, Yongbei Zhu, Liusu Wang, Hui Deng, et al. Mining whole-lung information by artificial intelligence for predicting egfr genotype and targeted therapy response in lung cancer: a multicohort study. *The Lancet Digital Health*, 4(5):e309–e319, 2022. 6

[50] Yan Wang, Yuyin Zhou, Wei Shen, Seyoun Park, Elliot K Fishman, and Alan L Yuille. Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. *Medical image analysis*, 55:88–102, 2019. 3

[51] Jakob Wasserthal, Manfred Meyer, Hanns-Christian Breit, Joshy Cyriac, Shan Yang, and Martin Segeroth. Totalsegmentator: robust segmentation of 104 anatomical structures in CT images. *arXiv preprint arXiv:2208.05868*, 2022. 3, 6

[52] WHO. Global health estimates 2020: Deaths by cause, age, sex, by country and by region, 2000–2019, 2020. 2

[53] Yingda Xia, Jiawen Yao, Le Lu, Lingyun Huang, Guotong Xie, Jing Xiao, Alan Yuille, Kai Cao, and Ling Zhang. Effective pancreatic cancer screening on non-contrast CT scans via anatomy-aware transformers. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 259–269. Springer, 2021. 3

[54] Yingda Xia, Qihang Yu, Linda Chu, Satomi Kawamoto, Seyoun Park, Fengze Liu, Jieneng Chen, Zhuotun Zhu, Bowen Li, Zongwei Zhou, et al. The felix project: Deep networks to detect pancreatic neoplasms. *medRxiv*, pages 2022–09, 2022. 3

[55] Yutong Xie, Yong Xia, Jianpeng Zhang, Yang Song, Dagan Feng, Michael Fulham, and Weidong Cai. Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest CT. *IEEE Transactions on Medical Imaging*, 38(4):991–1004, 2018. 3

[56] Yutong Xie, Jianpeng Zhang, Chunhua Shen, and Yong Xia. Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 171–180. Springer, 2021. 3

[57] Lianyan Xu, Ke Yan, Le Lu, Weihong Zhang, Xu Chen, Xiaofei Huo, and Jingjing Lu. External and internal validation of a computer assisted diagnostic model for detecting multi-organ mass lesions in CT images. *Chinese Medical Sciences Journal*, 36(3):210–217, 2021. 2

[58] Ke Yan, Jinzheng Cai, Youjing Zheng, Adam P. Harrison, Dakai Jin, Youbao Tang, Yuxing Tang, Lingyun Huang, Jing Xiao, and Le Lu. Learning from multiple datasets with heterogeneous and partial labels for universal lesion detection in ct. *IEEE Trans. Medical Imaging*, 40(10):2759–2770, 2021. 3, 6, 7, 8

[59] Ke Yan, Yifan Peng, Veit Sandfort, Mohammadhadi Bagheri, Zhiyong Lu, and Ronald M Summers. Holistic and comprehensive annotation of clinically significant findings on diverse CT images: learning from radiology reports and label ontology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8523–8532, 2019. 2, 3

[60] Ke Yan, Youbao Tang, Yifan Peng, Veit Sandfort, Mohammadhadi Bagheri, Zhiyong Lu, and Ronald M. Summers. Mulan: Multitask universal lesion analysis network for joint lesion detection, tagging, and segmentation. In *MICCAI*, 2019. 3

[61] Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of Medical Imaging*, 5(3):036501, 2018. 2, 3, 7

[62] Ke Yan, Xiaosong Wang, Le Lu, Ling Zhang, Adam P Harrison, Mohammadhadi Bagheri, and Ronald M Summers. Deep lesion graphs in the wild: relationship learning and organization of significant radiology image findings in a diverse large-scale lesion database. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, pages 9261–9270, 2018. 2, 3, 6, 7, 8

[63] Jiancheng Yang, Yi He, Kaiming Kuang, Zudi Lin, Hanspeter Pfister, and Bingbing Ni. Asymmetric 3d context fusion for universal lesion detection. In *MICCAI*, 2021. 3

[64] Xiaojun Yang, Lei Wu, Weitao Ye, Ke Zhao, Yingyi Wang, Weixiao Liu, Jiao Li, Hanxiao Li, Zaiyi Liu, and Changhong Liang. Deep learning signature based on staging ct for preoperative prediction of sentinel lymph node metastasis in breast cancer. *Academic Radiology*, 27(9):1226–1233, 2020. 6

[65] Jiawen Yao, Kai Cao, Yang Hou, Jian Zhou, Yingda Xia, Isabella Nogues, Qike Song, Hui Jiang, Xianghua Ye, Jianping Lu, et al. Deep learning for fully automated prediction of overall survival in patients undergoing resection for pancreatic cancer: A retrospective multicenter study. *Annals of Surgery*, pages 10–1097, 2022. 6

[66] Jiawen Yao, Xianghua Ye, Yingda Xia, Jian Zhou, Yu Shi, Ke Yan, Fang Wang, Lili Lin, Haogang Yu, Xian-Sheng Hua, et al. Effective opportunistic esophageal cancer screening using noncontrast CT imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 344–354. Springer, 2022. 3, 6

[67] Lisha Yao, Yingda Xia, Haochen Zhang, Jiawen Yao, Dakai Jin, Bingjiang Qiu, Yuan Zhang, Suyun Li, Yanting Liang, Xian-Sheng Hua, et al. Deepcrc: Colorectum and colorectal cancer segmentation in CT scans via deep colorectal coordinate transform. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 564–573. Springer, 2022. 3, 6

[68] Paul A. Yushkevich, Joseph Piven, Heather Cody Hazlett, Rachel Gimpel Smith, Sean Ho, James C. Gee, and Guido Gerig. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage*, 31(3):1116–1128, 2006. 6

[69] Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1195–1204, 2021. 3

[70] Yao Zhang, Jiawei Yang, Jiang Tian, Zhongchao Shi, Cheng Zhong, Yang Zhang, and Zhiqiang He. Modality-aware mutual learning for multi-modal medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 589–599. Springer, 2021. 3

[71] Tianyi Zhao, Kai Cao, Jiawen Yao, Isabella Nogues, Le Lu, Lingyun Huang, Jing Xiao, Zhaozheng Yin, and Ling Zhang. 3d graph anatomy geometry-integrated network for pancreatic mass segmentation, diagnosis, and quantitative patient management. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13743–13752, 2021. 2, 3, 6

[72] Sunyi Zheng, Jiapan Guo, Xiaonan Cui, Raymond NJ Veldhuis, Matthijs Oudkerk, and Peter MA Van Ooijen. Automatic pulmonary nodule detection in CT scans using convolutional neural networks based on maximum intensity projection. *IEEE Transactions on Medical Imaging*, 39(3):797–805, 2019. 3

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

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

[75] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3

[76] Zhuotun Zhu, Yingda Xia, Lingxi Xie, Elliot K Fishman, and Alan L Yuille. Multi-scale coarse-to-fine segmentation for screening pancreatic ductal adenocarcinoma. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–12. Springer, 2019. 2, 3

Appendix: CancerUniT

Abstract. This document contains the Supplementary Materials for the ICCV 2023 paper "CancerUniT: Towards a Single Unified Model for Effective Detection, Segmentation, and Diagnosis of Eight Major Cancers Using a Large Collection of CT Scans". It covers the model generalizability to public dataset, (§A), model instantiation details (§B), semantic segmentation results of full spectrum tumors (§C), and the qualitative results (§E).

A. Generalizability to Public Dataset

Our method aims at holistically modeling the multiple cancer screening problem versus non-cancer. However, to the best of our knowledge, no public dataset is suitable for such problems. Nevertheless, our trained model generalizes well on three public single-tumor datasets including MSD pancreas, liver and lung dataset, as shown in Table. A. It is worth noting that our model inference directly without extra training, whereas the 3 single-nnUNets is trained on the MSD dataset with domain knowledge. To be specific, the experiment of 3 single-nnUNet is conducted with 5-fold cross-validation, and our UniT is tested on the same validation set.

Despite not having any prior knowledge of the data distribution, our proposed UniT model effectively suppresses the single-tumor expert model, achieving an average tumor detection sensitivity improvement of 3.1%. Our results demonstrate the efficacy of our proposed method for addressing the tumor detection problem without the need for a specific dataset. The ability to generalize well on public datasets and suppress the single-tumor expert model underscores the potential of our approach to be used as a practical solution for universal cancer screening and diagnosis.

Table 8. Generalizability to 3 Public MSD dataset [2]. Average detection sensitivity is reported. Our model inference directly, whereas 3 single-nnUNets are trained on the MSD dataset.

	Pancreatic tumor	Liver tumor	Lung tumor	Avg	Speed	Param
single-nnUNet (trained)	88%	97%	90.5%	91.8%	66s	92.34M
Ours (test)	94.7%	93.1%	97%	94.9%	42s	30.87M

B. Model Instantiation Details

In our UniT, the hidden dimension of query is set to 32, such that the detection query $\mathbf{A}^j \in \mathbb{R}^{4\times32}$, the diagnosis query $\mathbf{B}^j \in \mathbb{R}^{10\times32}$, the shared query $\mathbf{S}^j \in \mathbb{R}^{12\times32}$. We adopt nnUNet [26] as the backbone to extract multi-scale features $\mathbf{F} = [\mathbf{F}^1, \mathbf{F}^2, \mathbf{F}^3, \mathbf{F}^4]$. Note, $\mathbf{F}^j \in \mathbb{R}^{d\times(D\times H\times W)}$ is flatten and projected from intermediate spatial feature $\hat{\mathbf{F}}^j \in \mathbb{R}^{C\times D\times H\times W}$. In specific, $\mathbf{F}^1 \in \mathbb{R}^{32\times(48\times192\times192)}$, $\mathbf{F}^2 \in \mathbb{R}^{32\times(48\times96\times96)}$, $\mathbf{F}^3 \in \mathbb{R}^{32\times(24\times48\times48)}$, and $\mathbf{F}^4 \in \mathbb{R}^{32\times(12\times24\times24)}$. The total number of Transformer layer is set to 3, each of which contains a multi-head cross-attention, a multi-head self-attention, and a feed-forward network. Note, in the inference stage, the tumor segmentation maps are extract to generate the tumor instances with class labels, where those tumor instances with less than 200 voxels are discarded.

C. Semantic Segmentation Results of Full Spectrum Tumors

We conducted an evaluation of our model's performance on the semantic segmentation of full spectrum tumors, which is a challenging task that involves the segmentation of multiple tumor subtypes within an organ. The quality of the multi-class tumor segmentation was assessed using the multi-class Dice score, where each subtype of the tumor was treated as an independent semantic class.

Our model outperformed the segmentation baselines and achieved the highest average segmentation Dice score, as demonstrated in Table 9. Notably, our model was not compared with detection models such as DeepLesion and LENS, as these models are not designed for semantic segmentation tasks.

Our findings suggest that enhancing the query hierarchy in our model can improve the semantic segmentation of full spectrum tumors. This observation is in line with our assumption that our query-based Transformer model can more effectively explore the similarity between intra-organ tumor subtypes, leading to improved segmentation performance. Overall, our evaluation provides evidence that our proposed model can effectively address the challenges of multi-class tumor segmentation in the context of full spectrum tumors.

D. Universal Cancer Screening: CT *vs* Blood Test.

Blood test is now one of the most attractive tools for non-invasive multi-organ cancer screening [13, 30, 28]. CT scanning had been considered historically for the same task, but was limited by its insufficient sensitivity and specificity [1]. AI reading in CT as an alternative opportunistic screening tool, our approach also has strong clinical potential for cancer detection screening. The advantage of CT is that this protocol is already an indispensable diagnostic imaging for cancer, but a positive blood test result requires further examinations for confirmation. With our model, clinicians have direct visual analyses of the detected cancer sites and misdetections of cancer can be largely reduced. No additional cost is needed under the opportunistic CT screening protocol whereas a single blood test can usually take $\sim$ 1000 US

Table 9. Voxel-level semantic segmentation results of full spectrum tumors. The Dice coefficient is reported. Note: the Dice values are calculated in a semantic manner, e.g., the HCC voxel is correctly segmented as the HCC subtype (not other liver tumor subtypes or other tumor types) by the semantic segmentation.

Model	Pancreas			Eso			Stomach			Liver					Average
	PDAC	nonPDAC	avg	EC	nonEC	avg	GC	nonGC	avg	HCC	ICC	Meta	Heman	avg	
8-nnUNet ensemble	0.750	0.525	0.638	0.770	0.433	0.602	0.441	0.099	0.270	0.489	0.552	0.296	0.784	0.530	0.510
nnUNet [26]	0.758	0.534	0.646	0.739	0.207	0.473	0.453	0.068	0.261	0.410	0.481	0.306	0.739	0.484	0.466
TransUNet [7]	0.749	0.553	0.651	0.744	0.321	0.533	0.473	0.128	0.301	0.411	0.503	0.353	0.717	0.496	0.495
Ours	0.728	0.560	0.644	0.738	0.457	0.597	0.389	0.187	0.288	0.368	0.666	0.305	0.773	0.528	0.514

dollars.

For relative performance comparison to CancerSeek [13], i.e., cancer vs. normal, our method has higher sensitivity levels in detecting six out of seven types of cancers: approximately for stomach (+18%), pancreas (+24%), esophagus (+26%), colorectum (+35%), lung (+34%), and breast (+57%). Our averaged patient-level cancer detection sensitivity is 94% versus 70% in [13]. The test specificity for normal cases in venous CT is 100% (blood test>99%). We acknowledge that the results of the representative blood test [13] and ours may not directly comparable since different test data are used. Nevertheless, the rough comparison shows the high accuracy of CT+AI solution, and thus may re-open doors for multi-cancer screening by CT [1].

E. Qualitative Results

We provide more qualitative results of full spectrum tumors in the test set being segmented and diagnosed by our method as shown in Fig. 4. The results demonstrate that our method can not only segment the tumor region well but also predict the class of tumor subtype correctly.

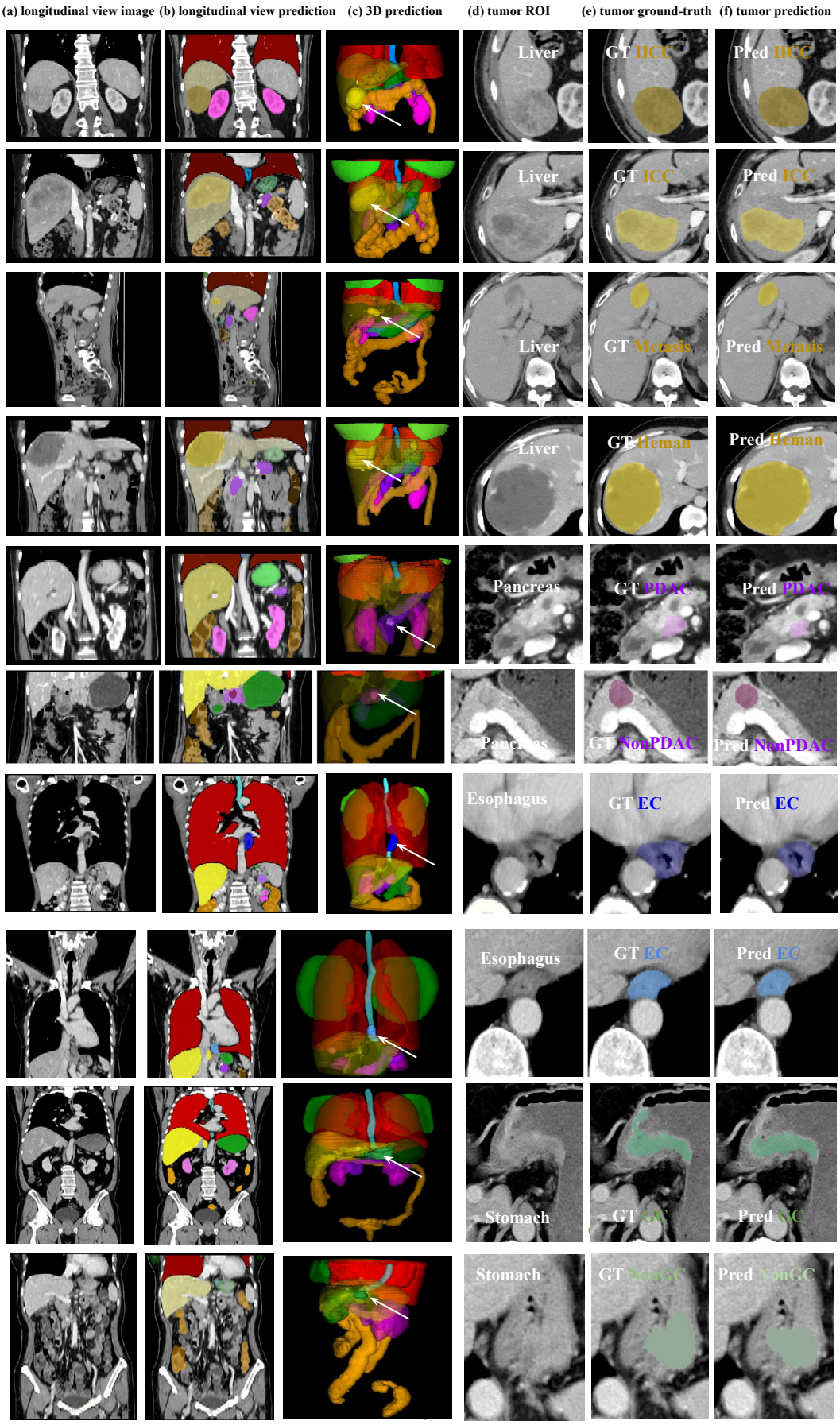

Figure 4. Qualitative results of full spectrum tumors in the test set being segmented and diagnosed by our method (best viewed in color).