Received: 6 June 2022 | Revised: 25 March 2023 | Accepted: 12 April 2023

DOI: 10.1002/mp.16438

RESEARCH ARTICLE

MEDICAL PHYSICS

Whole-body tumor segmentation from PET/CT images using a two-stage cascaded neural network with camouflaged object detection mechanisms

Jiangping He[1] | **Yangjie Zhang**[1] | **Maggie Chung**[2] | **Michael Wang**[3] | **Kun Wang**[1] | **Yan Ma**[1] | **Xiaoyang Ding**[1] | **Qiang Li**[1] | **Yonglin Pu**[4]

[1]Department of Electronic Engineering, Lanzhou University of Finance and Economics, Lanzhou, Gansu, China

[2]Department of Radiology, University of California, San Francisco, California, USA

[3]Department of Pathology, University of California, San Francisco, California, USA

[4]Department of Radiology, University of Chicago, Chicago, Illinois, USA

Correspondence
Yonglin Pu, Department of Radiology, University of Chicago, Chicago, IL 60637, USA.
Email: ypu@radiology.bsd.uchicago.edu

Funding information
Nature Science Foundation of Gansu province, China, Grant/Award Numbers: 20JR5RA200, 22ZD6GA047, 21JR11RA134; Nature Science Foundation of Lanzhou University of Finance and Economics, Grant/Award Number: Lzufe2021W-002

Abstract

Background: Whole-body Metabolic Tumor Volume (MTVwb) is an independent prognostic factor for overall survival in lung cancer patients. Automatic segmentation methods have been proposed for MTV calculation. Nevertheless, most of existing methods for patients with lung cancer only segment tumors in the thoracic region.

Purpose: In this paper, we present a Two-Stage cascaded neural network integrated with Camouflaged Object Detection mEchanisms (TS-Code-Net) for automatic segmenting tumors from whole-body PET/CT images.

Methods: Firstly, tumors are detected from the Maximum Intensity Projection (MIP) images of PET/CT scans, and tumors' approximate localizations along *z*-axis are identified. Secondly, the segmentations are performed on PET/CT slices that contain tumors identified by the first step. Camouflaged object detection mechanisms are utilized to distinguish the tumors from their surrounding regions that have similar Standard Uptake Values (SUV) and texture appearance. Finally, the TS-Code-Net is trained by minimizing the total loss that incorporates the segmentation accuracy loss and the class imbalance loss.

Results: The performance of the TS-Code-Net is tested on a whole-body PET/CT image data-set including 480 Non-Small Cell Lung Cancer (NSCLC) patients with five-fold cross-validation using image segmentation metrics. Our method achieves 0.70, 0.76, and 0.70, for Dice, Sensitivity and Precision, respectively, which demonstrates the superiority of the TS-Code-Net over several existing methods related to metastatic lung cancer segmentation from whole-body PET/CT images.

Conclusions: The proposed TS-Code-Net is effective for whole-body tumor segmentation of PET/CT images. Codes for TS-Code-Net are available at: https://github.com/zyj19/TS-Code-Net.

KEYWORDS
camouflaged object detection, tumor segmentation, two-stage cascaded neural network, whole-body PET/CT

1 | INTRODUCTION

Positron emission tomography (PET)/computed tomography (CT) imaging is a widely used multimodality imaging tool integral to prognostic assessment and treatment planning of cancer patients. 18F-Fluorodeoxyglucose positron emission tomography (FDG-PET) measures the amount of metabolic activity throughout the body using the radioisotope tracer fluoro-18-deoxyglucose (18F-FDG), a glucose analog. Cancerous tissues are hypermetabolic with increased glucose uptake. Compared to FDG-PET, CT offers high spatial resolution which is important for localization. Combined PET/CT allows the detection of hyper-metabolic tumors with corresponding anatomic information.

Lung cancer is the second most common cancer in the world, representing 11.4% of all cancers in 2020. It is also the leading cause of cancer related deaths worldwide, associated with 18.0% of total cancer related deaths.[1] In 2022, approximately 350 deaths per day from lung cancer are projected to occur in the United States.[2] Non-small cell lung cancer (NSCLC) accounts for about 80% to 85% of lung cancers.[3] Imaging evaluation is critical for predicting survival and guiding clinical management of lung cancer patients. Metabolic Tumor Volume (MTV) refers to the metabolically active volume of segmented tumor. Whole-body Metabolic Tumor Volume (MTVwb) has been shown to be an independent prognostic variable of Overall Survival (OS) in lung cancer patients.[4,5] The MTVwb can be measured on whole-body PET/CT images through manual segmentation by radiologists, however it is a time consuming and tedious task. Automatic segmentation of lung cancer tumors to calculate MTVwb can assist radiologists in their assessment of lung cancer patients.

Many automatic methods have been proposed for tumor segmentation in PET/CT images such as active shape model based methods,[6,7] texture feature and classifier based methods,[8,9] graph based methods,[10–12] and so on. A brief description of the different types methods can be found in the educational report.[13] Recently, deep learning based methods[14–18] have been widely used in tumor segmentation of PET/CT images for a variety of cancer types.[19–22] Huang et al. segmented head and neck tumors using Convolutional Neural Network (CNN) that contained a feature representation phase and score-map reconstruction phase.[23] Xu et al. used a W-Net[24] for whole-body bone tumor segmentation consisting of two V-Nets.[25] Blanc-Durand et al. performed lymphoma tumor segmentation from whole-body PET/CT by using a 3D U-Net[26] architecture with two input channels.[27] Revailler et al. utilized a 3D V-Net model to generate lymphoma tumor segmentations with soft dice loss from PET/CT images.[28]

Deep learning also has been utilized in lung cancer segmentation. Many methods proposed integration of information from PET and CT images. Zhong et al. used deep Fully Convolutional Networks (FCNs) to co-segment lung tumor from PET/CT images.[29] Li et al. used a 3D FCN to generate a probability map from CT images, which was then embedded into the PET images for tumor segmentation.[30] Kumar et al. trained a Co-learn neural network (Co-learn-Net) to fuse complementary information from thoracic PET/CT images. This network produces a fusion map that explicitly quantifies the fusion weights in different areas (lung, tumor, mediastinum and background) in each modality for tumor segmentation.[31] Fu et al. proposed a deep learning-based framework in multi-modal PET/CT segmentation with a Multi-modal Spatial Attention Module network (MSAM-Net). Their method automatically learns to emphasize regions related to tumors and suppresses normal regions with physiologic uptake on PET images.[32]

While these published methods for lung cancer appear promising in segmentation accuracy, their methods are not intended for quantifying distant metastatic disease outside of the thoracic region. Existing studies demonstrating the correlation between MTV and survival have used MTV from the whole body (MTVwb) for prognostication and treatment guidance,[4,5] and thus the thoracic-only MTV is not a sufficient approximation of the MTVwb. Therefore, a whole-body PET/CT tumor segmentation method is needed for accurate assessment of disease extent and prognostication.

Compared with thoracic tumor segmentation on PET/CT, there are several challenges associated with whole-body tumor segmentation:

1. Extra-thoracic tumors may have a similar tissue density as the surrounding soft tissue, making the tumor borders difficult to identify on CT images.
2. Normal organs and tissues can be metabolically active (FDG-avid), and therefore appear similar to tumors on PET images. Examples include the brain, spleen, liver, bone marrow, heart, lymphoid tissue, and brown fat.
3. FDG is renal excreted, thus the kidneys, ureters, and bladder may also appear FDG-avid.

Figure 1 is an example that demonstrates the challenges associated with the whole-body tumor segmentation from PET/CT images. A PET slice and its corresponding CT slice at the same position and the Maximum Intensity Projection (MIP)[33] image of the PET scan from the font view show that FDG-avid tumors are difficult to distinguish from normal anatomy.

In this study, we treat tumor segmentation from whole-body PET/CT images as cascade object detection and segmentation problems. Hence, we present a two-stage cascaded neural network integrated with camouflaged

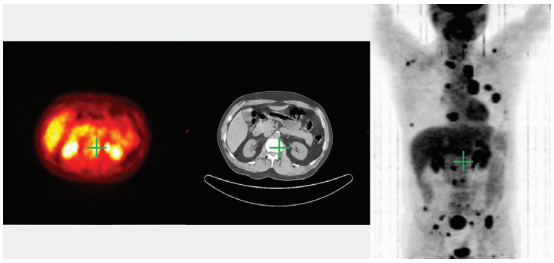

FIGURE 1 A patient's whole-body PET/CT scan. A PET slice with heat scores (left) and its corresponding CT slice (middle), and the MIP image (right). The green cross indicates one of tumors of the patient.

object detection mechanisms[34,35] for tumor segmentation from whole-body PET/CT images.

The contributions of our method are summarized below:

1. a two-stage architecture is designed to separate the complex task of tumor segmentation from whole body into simpler tasks of tumor detection and tumor segmentation;
2. tumors are detected from the MIP images which benefits from global contextual information of the entire body;
3. Camouflaged object detection mechanisms are utilized to identify tumors that appear similar to the surrounding anatomy.

We compare our experimental results of the whole-body PET/CT tumor segmentation with several existing methods on a data-set of 480 NSCLC patients.

2 | METHOD

We present a Two-Stage neural network integrated with Camouflaged Object Detection mechanisms (TS-Code-Net) for tumor segmentation from whole-body PET/CT images. A camouflaged object in our case is a tumor that appears similar to the surrounding normal anatomy. The architecture of our method consists of two stages. First, the approximate tumors locations are obtained from the MIP images of both PET and CT scans, then the segmentations are performed using the tumor containing slices detected by the first step. The loss function of TS-Code-Net contains two terms, which are Dice loss $\mathcal{L}_{dice}$ to evaluate the segmentation accuracy and the focal loss $\mathcal{L}_{focal}$ to reduce the class-imbalance effect. The details of our method are described below.

2.1 | Structure of the TS-code-net

There are two stages in our method: tumor localization (Stage I) from the MIPs of PET and CT images, and tumor segmentation (Stage II) from the slices containing tumors detected at Stage I. Both stages have similar architectures. Figure 2 shows the work-flow of our method.

In stage I, MIP images are separately obtained from PET and CT images by the MIP[33] from the front view. Both MIP images are then fed to separate encoders for feature extraction. The features extracted from the two modalities are fused together to take advantage of the complementarity of PET and CT scans, followed by a Feature Enhancement module to increase feature discriminability. After that, the fused and enhanced features are transferred to a decoder to generate tumor probability maps for each patient. Finally, 3D patches of PET and CT containing detected tumors are identified, as shown in Stage I of Figure 2.

Voxels with high tumor probability are found by comparing the probability values of neighboring voxels. Their positions at the z-axis of the PET and CT images are regarded as the locations of tumor-containing slices. Subsequently, 3D patches are formed by a pre-specified number (32 in our experiments) of slices below and above the voxels along the z-axis.

The voxel's position at the z-axis can be located from the 2D tumor probability map generated by 2D MIP images. This is because the height of the MIP images corresponds to the z-axis of the 3D PET/CT images. In a 2D tumor probability map, there might be more than two pixels with high tumor probability in a tumor region. However, we need only one pixel's position to locate the tumor and form the 3D patches for the next stage. The Non-Maximum Suppression operation[36] is then performed to filter out the best pixel by a pre-specified size sliding window (say, 48×48). The NMS algorithm will accept the pixel with highest score and reject its neighboring pixels located in the same tumor area, as it does in the computer vision tasks, such as object detection and interesting points detection.

Stage II contains a similar structure to stage I, with 3D patches, 3D encoder, and 3D decoder instead of 2D counterparts. Similar procedures described in Stage I are applied to 3D PET and CT patches. 3D tumor segmentation is obtained after the 3D decoder.

2.2 | The encoders and decoders in our method

The encoders of stage I and stage II have the same architecture. 2D convolution kernels are used in stage I and 3D convolution kernels are used in stage II. The decoders of stage I and stage II have similar architectures with few differences. For the sake of brevity, we only describe the encoder and decoder architectures of stage II as below (see Figure 3).

In Figure 3, "UP" is an up sampling operation. "UP× n" means n up sampling operations are applied. The 3D

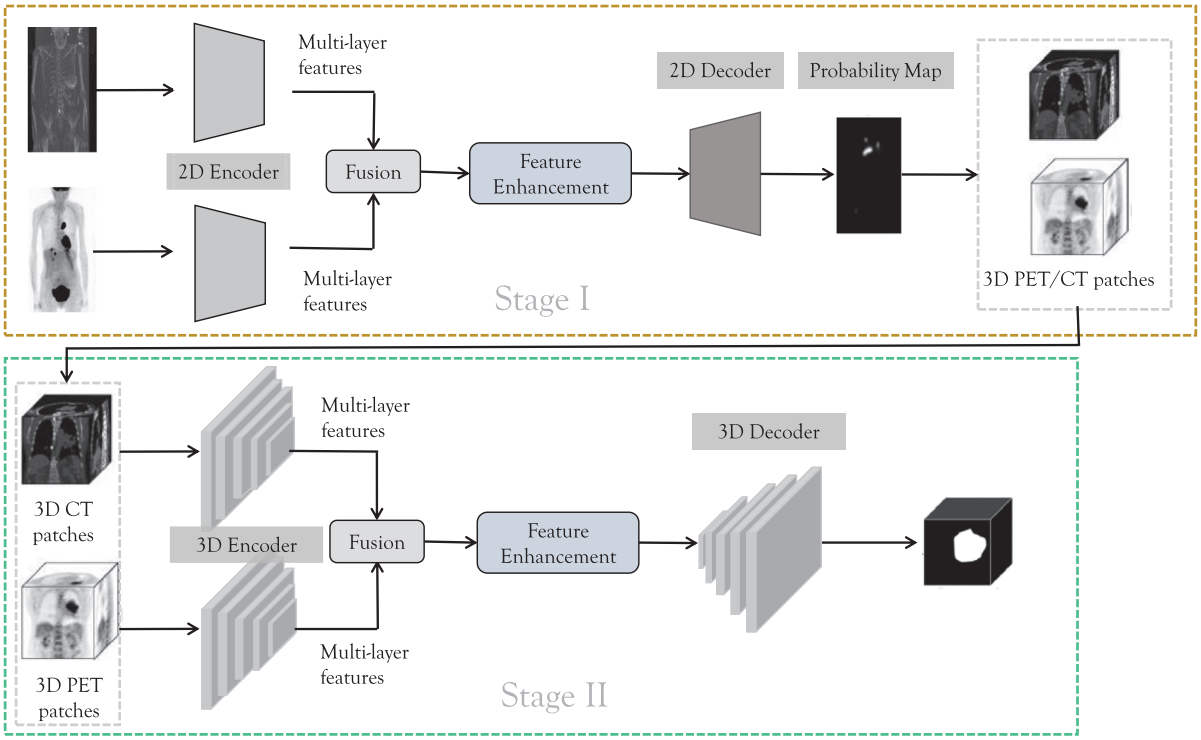

FIGURE 2 The overall flowchart of the TS-Code-Net.

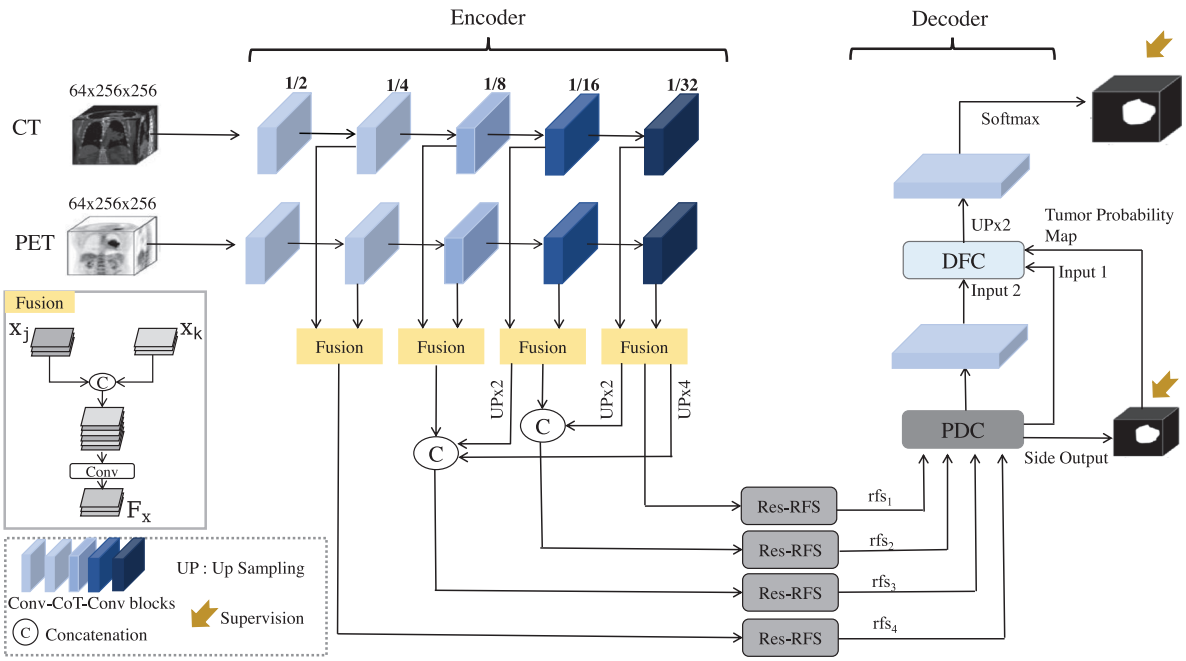

FIGURE 3 The 3D encoder and decoder in the TS-Code-Net.

PET and CT patches are fed to the encoder of stage II. In our study, the CoTNet[37] is adopted as the backbone network for the encoder since its transformer-style design can strengthen the capacity of visual representation. "Conv-CoT-Conv" refers to the combination of the convolution operation and the CoT block which is proposed in the CoTNet. After multiple down sampling layers, the multi-layer PET and CT features are fused using the method described in Co-learn-Net[31] by the following two steps. First, a spatially varying fusion

map is derived based on encoded modality-specific features. Second, the modality-specific features and the fusion maps of PET and CT scans are multiplied to obtain a location-related representation of the complementary PET and CT information. The fused feature maps are concatenated after up sampling layers.

The Res-RFs (Receptive Fields[38]) are applied to the concatenated features to enhance discriminative representations by combining multiple-field features obtained by different kernel-size and dilation-rate convolution layers. The flowchart of the Res-RFs is provided in the appendix of our paper. Please note that we modify the original RFs by adding skip-connects, a similar structure in the ResNet,[17] and the Context Exploration(CE) block[35] for better performance (the details of the CE block are provided in the Decoding Focusing Component [DFC] subsection).

The feature maps rfs_i are then fused by the Partial Decoder Component (PDC) module,[34] which has an excellent capability for different levels feature fusions. More specifically, different levels of features rfs_i are integrated using multiple up sample operations and the Bconv operations. Bconv stands for a sequential operation that combines a 3×3 convolution, a batch normalization and a ReLU function. Thus, a coarse camouflage map can be computed. The flowchart of PDC and its details can be found in Fan et al.'s paper.[34]

Tumors may have similar appearances compared to their surroundings in PET/CT images, especially in metabolically active (FDG-avid) areas. The PDC module can produce coarse camouflage maps of tumors. However, it is still not easy to identify the tumors. To obtain precise segmentations, we first add a side output to the PDC module with the supervision derived from the ground truth to guide the segmentation. Then, we modify the PDC module into a DFC module and apply both of them to distinguish the tumors and their surrounding regions using the camouflaged object detection mechanisms. Finally, the DFC's output generates tumor segmentation results after up sampling layers and a softmax activation.

The original PDC has only one output, that is, the tumor segmentation map. In our design, we let the PDC module has three same outputs by the duplication operation since the DFC module is adopted. The first output is led to the subsequent convolution layers for semantic information extraction. The second output is connected to the following DFC module to function as a skip connection. The third output is compared to the ground truth forming a side output following a $1 \times 1 \times 1$ convolution and a sigmoid activation.[39] The side output generates the tumor probability maps by computing the side losses with ground truth supervision to the benefit of locating the tumors in the followed DFC module.

2.3 | DFC

PDC is used to achieve multi-scale feature fusion for its ability to integrate different levels of features. We extend it to the DFC module by adding a Focus Module (FM)[35] for obtaining more accurate tumor contours. Therefore, tumors and their surroundings can be more precisely identified. The architectures of the DFC module is shown in Figure 4.

In the DFC module, BConvLR is similar to BConv of the PDC module but with a Leaky ReLU[40] function instead of the ReLU function. Input 1 is features with low levels such as gray scale, edges obtained from low layers and Input 2 is features with high levels such as parts, objects obtained from high layers. The PDCseg is the tumor segmentation map produced by the side output of the PDC module. The up sampling layers and BConvLR are first applied to fuse the three inputs, Input 1, Input 2, and PDCseg of DFC, as described in Figure 4. Due to the similarities that some tumors share with their surroundings, both false positive and false negative segmentations would be expected. We then utilize the FM[35] to let the network focus on the ambiguous regions within the unrefined tumor segmentation map PDCseg. After that, the maps are concatenated together and further processed by multiple BConvLR operations to provide a more precise tumors map.

Many tumors in our database can be regarded as camouflaged objects due to their similar appearance to their surroundings. The FM can first discover and then remove the false predictions including the false-positive distractions and the false-negative distractions.

In the stage of the distraction discovery, the PDCseg and its reverse version, $1 - PDCseg$, are multiplied to the low level features to generate the foreground-attentive features F_a and the background-attentive features F_b, respectively. Then the two types of features are fed into two CE blocks[35] to perform contextual reasoning. Both false-positive distractions and false-negative distractions can be found through the reasoning.

The CE block consists of four branches and each branch includes two convolutions with different kernel sizes. One is for channel reduction and the other is for local feature extraction, followed by a dilated convolution for context perceiving. The outputs of the four branches are finally concatenated and fused together for the capability of obtaining much more context information over a wide range of scales. Finally, the CE block can be applied to context reasoning and distraction discovery.

In the stage of the distraction suppression, the false-positive distractions will be removed from the high level feature maps and the false-negative distractions will be added to the high level feature maps, which is formulated as follow:

$$F_{out} = UP(F_H) - \alpha CE(F_a) + \beta CE(F_b) \quad (1)$$

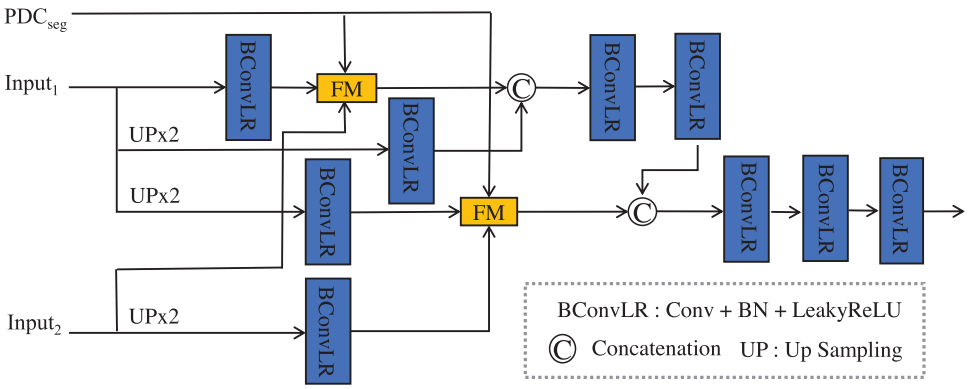

FIGURE 4 The proposed DFC module. DFC, Decoding focusing component.

where UP means up sampling and CE means the context exploration operation. F_H denotes the high level feature maps. α and β are two learnable scale parameters and they are initially set to 1. The fine tumors segmentation are then obtained after the distractions removal. Further details on FM and the CE block can be found in Mei et al.'s paper regarding camouflaged object detection.[35]

2.4 | Loss function

In our whole-body scan data-set, some tumors are very small compared to FDA-avid organs or larger tumors. These small tumors are often ignored by algorithms due to their comparative small contributions to the loss calculation. A combination of the segmentation loss (dice score) and the imbalance loss (focal loss score) was used to alleviate the small-object segmentation problem.[41] Thus, we adopt the combined loss in our study and the total loss function of our method is described in Equation (2).

$$\begin{aligned}\mathcal{L}_{total} &= \mathcal{L}_{Dice} + \lambda \times \mathcal{L}_{Focal} \\ &= C - \sum_{i=0}^{C-1} \frac{TP_p(i)}{TP_p(i) + aFN_p(i) + bFP_p(i)} \\ &\quad - \lambda \frac{1}{N} \sum_{i=0}^{C-1} \sum_{n=1}^{N} g_n(i)(1 - p_n(i))^2 \log(p_n(i)) \end{aligned} \quad (2)$$

In Equation (2), C denotes the class number ($C = 2$ here for the two classes of background and foreground). N is the total number of voxels to be processed. $TP_p(i)$, $FN_p(i)$ and $FP_p(i)$ are the probabilities of the true positives, false negatives and false positives for class i, respectively. The parameters a, and b are the balance weights for $FN_p(i)$ and $FP_p(i)$. $p_n(i)$ is the predicted probability for the voxel n which belongs to the class i; $g_n(i)$ is the ground truth for that voxel; λ is the trade-off between the dice loss and the focal loss.

3 | EXPERIMENTS

3.1 | Data

Our data-set includes de-identified whole-body PET/CT images from 480 Institutional Review Board approved NSCLC patients, retrospectively collected at the University of Chicago between 2004 and 2014. Three hundred fifty-two PET/CT scans are from a Reveal HD scanner (CTI, Knoxville, TN, USA) before 15 March 2012, and 128 PET/CT scans are from a Siemens CT scanner (Biograph mCT) on and after 15 March 2012.[5] The PET and the CT slices of each patient are registered by affine registration and have a final resolution of 256 × 256 pixels.

The voxel values of original PET volumes are converted to Standard Uptake Values (SUV) for normalization between patients. The PET/CT scans have different numbers of slices ranging from 189 to 543, and the number of tumors in each patient varies between 1 and 57. The ground truth for tumors segmentation are delineated from PET scans and provided by two board certified radiologists using the software MIMvista 5.1.2. The statics concerning patients' attributes are described in Table 1.

3.2 | Implementation details

We use a five-fold cross validation on our experiments. All 480 patients are randomly shuffled, and for each training-testing loop 96 patients are used for testing and the remaining patients are used for training. The whole procedure is repeated five times. Our experimental environment is constructed using Python 3.8 and TensorFlow 2.4. The graphical processing unit is

TABLE 1 The Characteristics of Study Sample.

Overall:	480(100%)	Gender, *N*(%)	
Age: categorical, *N*(%)		Female	267(55.6%)
34 and younger	1(0.2%)	Histology, *N*(%)	
35–44	6(1.3%)	Adeno	244(50.8%)
45–54	51(10.6%)	Squamous	144(30.0%)
55–64	120(25.0%)	Large cell	21(4.4%)
65 and older	302(62.9%)	NSCLC(NOS)	61(12.7%)
Race, *N*(%)		Other	10(2.1%)
White	186(38.8%)	Clinical TNM stage, *N*(%)	
Black	280(58.3%)	I	116(24.2%)
Other	14(2.9%)	II	59(12.3%)
Smoking status, *N*(%)		III	153(31.9%)
Never	39(8.1%)	IV	152(31.7%)
Current	170(35.4%)	Stratum MTVwb, *N*(%)	
Prior	271(56.5%)	I	131(27.3%)
Treatment category, *N*(%)		II	129(26.9%)
With surgery	156(32.5%)	III	112(23.3%)
No surgery	273(56.9%)	IV	108(22.5%)
No treatment	51(10.6%)		
Survival time, median (interquartile range) (day): 652.5(264.5–1821)			

Abbreviations: MTVwb, whole-body metabolic tumor volume; NSCLC, non-small cell lung cancer.

NVIDIA GeForce 3090 with 24 GB memory. In the training phase, we set the learning rate to be 0.001. The batch size of our method is 3 and the optimizer is NAdam. The threshold applied to the segmentation probability map is 0.4. The value of the λ in Equation (2) is 10. Codes for the TS-Code-Net are on: https://github.com/zyj19/TS-Code-Net.

3.3 | Implementation of comparison methods

We compare our method with five tumor segmentation methods for whole-body tumor segmentation using PET/CT images: Co-learn-Net[31] and MSAM-Net[32] that are specific for thoracic PET/CT lung cancer tumor segmentation; the PET/CT lymphoma segmentation method for whole-body tumor segmentation via V-Net[28]; the Densely Connected convolutional neural network (DC-Net) for automated segmentation of adipose tissue using whole-body MRI[42]; and the Seg-Net[43] a general segmentation method.

3.3.1 | Co-learn-Net

We re-implement the Co-learn-Net based on the published codes[44] without lung and mediastinum segmentation prior to training.

3.3.2 | MSAM-Net

We re-implement the MSAM-Net based on the original paper[32] and experimentally select best parameters for training.

3.3.3 | Whole-body lymphoma V-Net

Compared to the lymphoma segmentation method based on the V-Net,[28] we set 16 channels in the network instead of 8 to have improved segmentation performance.

3.3.4 | DC-Net

We re-implement the DC-Net based on the previously published codes.[42] The input size and other parameters are modified to adapt to PET/CT tumor segmentation.

3.3.5 | Seg-Net

Seg-Net was designed for single modality imaging tasks.[43] We concatenate PET and CT images in the channel axis to implement a dual modality input.

In the implementations of comparison methods, the optimizer for all of the above methods are NAdam, with a learning rate of 0.001. The batch size of Co-lean-Net and MSAM-Net are 16, of the whole-body lymphoma V-Net is 3, of DC-Net is 2 and of Seg-Net is 32 due to GPU memory limitations.

3.4 | Results

3.4.1 | Metrics

The Dice, Sensitivity, and Precision are applied to evaluate the performance of all methods, as shown in Equations (3)–(5):

$$Dice = \frac{2 \times |V_P \cap V_G|}{|V_P| + |V_G|} \tag{3}$$

$$Sensitivity = \frac{|V_P \cap V_G|}{|V_G|} \tag{4}$$

$$Precision = \frac{|V_P \cap V_G|}{|V_P|} \tag{5}$$

where V_P denotes the predicted volume, and V_G denotes the ground-truth volume. $|V|$ denotes the nonzero voxel number in volume V. If $|V_P|$ and $|V_G|$ of a patient are both0, the corresponding metrics is defined

TABLE 2 TS-Code-Net's comparison with other methods (mean ± std.).

	Dice	Precision	Sensitivity
Co-learn-Net	0.53±0.05	0.52 ±0.07	0.71 ± 0.03
MSAM-Net	0.55 ± 0.02	0.58±0.04	0.66 ±0.02
Whole-body lymphoma V-Net	0.50±0.02	0.61 ±0.06	0.54 ± 0.03
DC-Net	0.54 ±0.04	0.59 ± 0.02	0.63 ±0.07
Seg-Net	0.47 ±0.03	0.60 ±0.02	0.48 ±0.07
Ours	**0.70 ±0.03**	**0.70 ±0.04**	**0.76 ±0.04**

as 1 to avoid the zero-division error. If only one of the two values is 0, then the result is 0.

3.4.2 | Comparison and analysis

All metrics are calculated for each patient independently, the means and standard deviations of the performance measures are calculated over the five-fold cross-validation for comparison.

Comparison with other methods

We compare our method with other current methods using whole-body PET/CT images from the same training and testing patients. As shown in Table 2, the TS-Code-Net achieves highest scores in Dice, Sensitivity, and Precision. The MSAM-Net has the second highest Dice score, and it is 0.15 lower than our method. Dice scores of most methods are between 0.47 and 0.55, while our method demonstrates a 0.70 Dice score. The Sensitivity and Precision results are also superior using our method. These metrics support that the TS-Code-Net is more accurate for whole-body tumor segmentation compared to other related methods.

MIP of the PET volume is an overview for areas of uptake in the entire body. Figure 5 shows three MIPs with tumor contours segmented by radiologists, our method and the comparison methods.

In Figure 5, the left patient has only one tumor in the left hilar region; the middle one has multiple tumoral lesions in the thoracic and right supraclavicular region; the right one has extensive tumor lesions in the whole-body. The images show that our segmentation results are closer to the ground truth than those with other methods, especially for the tumor segmentation outside thoracic. The original images are cropped for a better demonstration and the whole body MIP images can be found in the appendix (see Figures A3 and A4).

Ablation study for the TS-Code-Net

To further verify the effectiveness of core structures, we conduct ablative experiments with four different settings, as shown in Table 3. Setting 1: the entire TS-Code-Net; Setting 2: the TS-Code-Net without the tumor localization stage, which means 3D PET/CT images are fed to the tumor segmentation network directly; Setting 3: the TS-Code-Net without the PDC and the DFC modules, which means no camouflaged object detection mechanisms are used; Setting 4: the TS-Code-Net without the DFC module but the PDC remained, which means no side output and the FM are used.

Table 3 shows that the removal of the tumor localization stage from TS-Code-Net causes the Dice to decrease by 0.07 (Setting 2). The camouflaged object detection mechanisms are useful for whole body tumor segmentation and the dice is only 0.66 if the mechanisms are not applied (Setting 3). When the DFC module is removed, the Dice decreased by 0.02 (Setting 4). This ablative experiment proves that all the components in the proposed TS-Code-Net are essential to the overall performance.

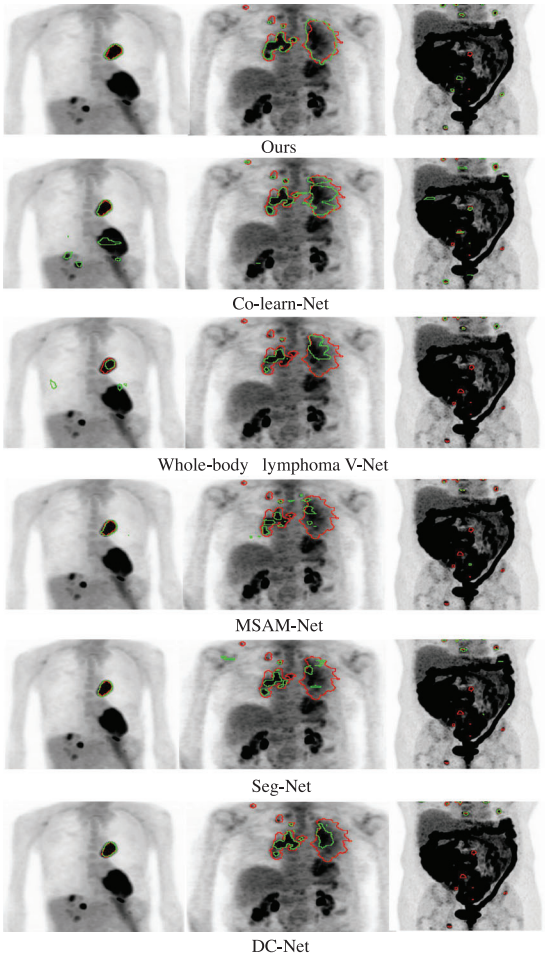

FIGURE 5 Three sample patients' MIPs with tumor contours delineated by radiologist (red) and different methods (green).The original images are cropped for a better demonstration.MIPs, Maximum intensity projections.

3.5 | Discussion

As previously discussed, the whole body contains many FDG-avid organs and tissues that share a similar

TABLE 3 Ablation study of our TS-Code-Net (mean ± std.).

	Dice	Precision	Sensitivity
1: TS-Code-Net	**0.70 ±0.03**	**0.70 ±0.04**	**0.76 ±0.04**
2: Without the detection stage	0.63 ±0.03	0.61 ±0.04	0.77 ±0.02
3: Without the PDC and DFC modules (No camouflaged object detection units)	0.66 ±0.03	0.67 ±0.05	0.73 ±0.05
4: Without the DFC module	0.68 ±0.03	0.69 ±0.05	0.76 ±0.04

Abbreviations: DFC, decoding focusing component; PDC, partial decoder component.

intensity as the tumors of interest. For this reason, whole-body segmentation for FDG-avid tumors is more difficult than thoracic-only segmentation. Compared to other cited methods, our method achieves better tumor segmentation results, verifying the effectiveness of our algorithm design.

Most methods we compared have available codes. Based on the published codes, we modified them and transferred them to be suitable to our experiments. Some methods performed not well as described in their original papers, and we believe that the gaps are from inherent differences between our whole-body PET/CT data-set and other data sets the comparison methods used.

In the whole-body PET/CT data-set, background voxels greatly outnumber tumor voxels. Our cascaded neural networks separate the complex whole body tumor segmentation into detection and segmentation problems. Only slices that contain tumor are used for segmentation, allowing higher memory efficiency and better segmentation accuracy. The ablation study shows a decrease in Dice score from 0.70 to 0.63 when only the tumor segmentation stage is used in the TS-Code-Net, showing that the cascade design is efficient for whole-body tumor segmentation.

Another obstacle of whole body tumor segmentation in PET/CT images is that tumors and normal organs and tissues can have similar FDG uptake. Thus reducing false negative and false positive segmentation can be difficult. The ablation setting 3 experiment left out the PDC and the DFC modules containing the camouflaged object detection and side output operations, which resulted in a decrease from 0.70 to 0.66 in the Dice score. This supports our hypothesis that camouflaged object detection benefits our whole-body tumor segmentation task, and its inclusion optimizes tumor segmentation accuracy.

Figure 5 and the figures in the appendix verify the advantages of the TS-Code-Net including more accurate segmentation of small tumors, and less over-segmentation of normal organs and tissues.

Our study has two main limitations. First, the data-set was retrospectively collected only for the NSCLC study and the patients with primary tumors of other cancers were excluded from this study. Second, the data-set was from the hospital of the University of Chicago, which means multi-center tests are not performed. In the future, more data will be collected with other hospitals' collaboration to address the limitations.

4 | CONCLUSION

In this paper, we propose the TS-Code-Net for whole-body tumor segmentation from PET/CT images of lung cancer patients. The TS-Code-Net contains two stages, which are tumor localization and tumor segmentation. Our method detects tumors from MIP images before the tumor segmentation stage, which reduces the burden of the segmentation network by discarding slices that contain no tumors. Camouflaged object detection mechanisms are also utilized to identify tumors that appear similar to their surroundings. By five-fold cross-validation, our method achieves 0.70, 0.76, and 0.70 on Dice, Sensitivity and Precision, which outperforms several existing methods related to tumor segmentation from PET/CT images.

ACKNOWLEDGMENTS

This work is supported by the Natural Science Foundation of Gansu Province, China (No. 20JR5RA200, No. 21JR11RA134, No. 22ZD6GA047) and the Natural Science Foundation of Lanzhou University of Finance and Economics (No. Lzufe2021W-002).

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts to disclose.

REFERENCES

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71:209-249.
2. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. *CA Cancer J Clin*. 2021;71:7-33.
3. American cancer society. What is non-small cell lung cancer? https://www.cancer.org/cancer/lung-cancer/about/what-is.html
4. Lee P, Weerasuriya DK, Lavori PW, et al. Metabolic tumor burden predicts for disease progression and death in lung cancer. *Int J Radiat Oncol Biol Phys*. 2007;69:328-333.
5. Zhang C, Liao C, Penney BC, Appelbaum DE, Simon CA, Pu Y. Relationship between overall survival of patients with non-small cell lung cancer and whole-body metabolic tumor burden seen on postsurgical fluorodeoxyglucose PET images. *Radiology*. 2015;275:862-869.

6. El Naqa I, Yang D, Apte A, et al. Concurrent multimodality image segmentation by active contours for radiotherapy treatment planning. *Med Phys*. 2007;34:4738-4749.
7. Markel D, Zaidi H, El Naqa I. Novel multimodality segmentation using level sets and Jensen-Rényi divergence. *Med Phys*. 2013;40:121 908.
8. Yu H, Caldwell C, Mah K, Mozeg D. Coregistered FDG PET/CT-based textural characterization of head and neck cancer for radiation treatment planning. *IEEE Trans Med Imaging*. 2009;28:374-383.
9. Watabe H, Markel D, Caldwell C, et al. Automatic segmentation of lung carcinoma using 3D texture features in 18-FDG PET/CT. *Int J Mol Imaging*. 2013;2013:13.
10. Song Q, Bai J, Han D, et al. Optimal co-segmentation of tumor in PET-CT images with context information. *IEEE Trans Med Imaging*. 2013;32:1685-1697.
11. Cui H, Wang X, Lin W, et al. Primary lung tumor segmentation from PET-CT volumes with spatial-topological constraint. *Int J Comput Assist Radiol Surg*. 2016;11:19-29.
12. Han D, Bayouth J, Song Q, et al.Globally optimal tumor segmentation in PET-CT images: A graph-based co-segmentation method. *Inf Process Med Imaging*. 2011:245-256.
13. Hatt M, Lee JA, Schmidtlein CR, et al. Classification and evaluation strategies of auto-segmentation approaches for PET: Report of AAPM task group no. 211. *Med Phys*. 2017;44:e1-e42.
14. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. *IEEE Conference on Computer Vision and Pattern Recognition*. 2017:2261-2269.
15. Lin TY, Piotr D, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE; 2017:936-944.
16. Girshick R. Fast R-CNN. *IEEE International Conference on Computer Vision*. IEEE; 2015:1440-1448.
17. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition*. 2016:770-778.
18. Gao SH, Cheng MM, Zhao K, Zhang XY, Yang MH, Torr P. Res2net: A new multi-scale backbone architecture. *IEEE Trans Pattern Anal Mach Intell*. 2021;43:652-662.
19. Oreiller V, Andrearczyk V, Jreige M, et al. Head and neck tumor segmentation in PET/CT: the HECKTOR challenge. *Med Image Anal*. 2022;77:102 336.
20. Shiri I, Vafaei Sadr A, Amini M, et al. Decentralized distributed multi-institutional PET image segmentation using a federated deep learning framework. *Clin Nucl Med*. 2022;47:606-617.
21. Jin D, Guo D, Ho TY, et al. Accurate esophageal gross tumor volume segmentation in PET/CT using two-stream chained 3D deep network fusion. In: Shen D, Liu T, Peters TM, Staib LH, Essert C, Zhou S, Yap PT, Khan A, eds. *Medical Image Computing and Computer Assisted Intervention-MICCAI 2019*, Springer International Publishing; 2019:182-191.
22. Chen L, Shen C, Li S, et al. Automatic PET cervical tumor segmentation by deep learning with prior information. Medical Imaging 2018: Image Processing. Vol. 10574. SPIE; 2018.
23. Huang B, Chen Z, Wu PM, et al. Fully automated delineation of gross tumor volume for head and neck cancer on PET-CT using deep learning: A dual-center study. *Contrast Media Mol Imaging*. 2018;2018:1-12.
24. Xu L, Tetteh G, Lipkova J, et al. Automated whole-body bone lesion detection for multiple myeloma on ^{68}ga-pentixa for PET/CT imaging using deep learning methods. *Contrast Media Mol Imaging*. 2018;2018:1-11.
25. Milletari F, Navab N, Ahmadi SA. V-Net: fully convolutional neural networks for volumetric medical image segmentation. *International Conference on 3D Vision*. 2016:565-571.
26. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. *Med Image Comput Comput Assist Interv*. 2015:234-241.
27. Blanc-Durand P, Jégou S, Kanoun S, et al. Fully automatic segmentation of diffuse large b cell lymphoma lesions on 3D FDG-PET/CT for total metabolic tumour volume prediction using a convolutional neural network. *Eur J Nucl Med Mol Imaging*. 2021;48:1-9.
28. Revailler W, Cottereau A, Rossi C, et al. Deep learning approach to automatize TMTV calculations regardless of segmentation methodology for major FDG-avid lymphomas. *Diagnostics*. 2022;12:417.
29. Zhong Z, Kim Y, Plichta K, et al. Simultaneous co-segmentation of tumors in PET-CT images using deep fully convolutional networks. *Med Phys*. 2019;46:619-633.
30. Li L, Zhao X, Lu W, Tan S. Deep learning for variational multimodality tumor segmentation in PET/CT. *Neurocomputing*. 2020;392:277-295.
31. Kumar A, Fulham M, Feng D, Kim J. Co-learning feature fusion maps from PET-CT images of lung cancer. *IEEE Trans Med Imaging*. 2020;39:204-217.
32. Fu X, Bi L, Kumar A, Fulham M, Kim J. Multimodal spatial attention module for targeting multimodal PET/CT lung tumor segmentation. *IEEE J Biomed Health Inform*. 2021;25:3507-3516.
33. Cody DD. AAPM/RSNA physics tutorial for residents: Topics in CTt. *RadioGraphics*. 2002;22:1255-1268.
34. Fan DP, Ji GP, Sun G, Cheng MM, Shen J, Shao L. Camouflaged object detection. *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE/CVF; 2020:2774-2784.
35. Mei H, Ji GP, Wei Z, Yang X, Wei X, Fan DP. Camouflaged object segmentation with distraction mining. *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE/CVF; 2021:8768-8777.
36. Paper with code. non maximum suppression, https://paperswithcode.com/method/non-maximum-suppression
37. Li Y, Yao T, Pan Y, Mei T. Contextual transformer networks for visual recognition. *IEEE Trans Pattern Anal Mach Intell*. 2023;45:1489-1500.
38. Liu S, Huang D, Wang Y. Receptive field block net for accurate and fast object detection. *European Conference on Computer Vision*. 2018:385-400.
39. Han J, Moraga C. The influence of the sigmoid function parameters on the speed of backpropagation learning. *Proceedings of the International Workshop on Artificial Neural Networks: From Natural to Artificial Neural Computation*. 1995:195-201.
40. Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks. *International Conference on Artificial Intelligence and Statistics*. 2011;15:315-323.
41. Zhu W, Huang Y, Zeng L, et al. AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Med Phys*. 2019;46:576-589.
42. Küstner T, Hepp T, Fischer M, et al. Fully automated and standardized segmentation of adipose tissue compartments via deep learning in 3D whole-body MRI of epidemiologic cohort studies. *Radiology: Artificial Intelligence*. 2020;2:e200 010.
43. Badrinarayanan V, Kendall A, Cipolla R. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans Pattern Anal Mach Intell*. 2017;39:2481-2495.
44. https://github.com/ashnilkumar/colearn

How to cite this article: He J, Zhang Y, Chung M, et al. Whole-body tumor segmentation from PET/CT images using a two-stage cascaded neural network with camouflaged object detection mechanisms. *Med Phys*. 2023;1-12. https://doi.org/10.1002/mp.16438

APPENDIX A

The Res-RF block is shown in Figure A1. Please note that we modify the original RFs by adding skip-connects and the Context Exploration block for better performance.

Illustrations of the outputs of the PDC and the DFC modules (see Figure A2).

The last two figures demonstrate tumor segmentation results using entire MIP images from the front view and the side view (see Figures A3 and A4).

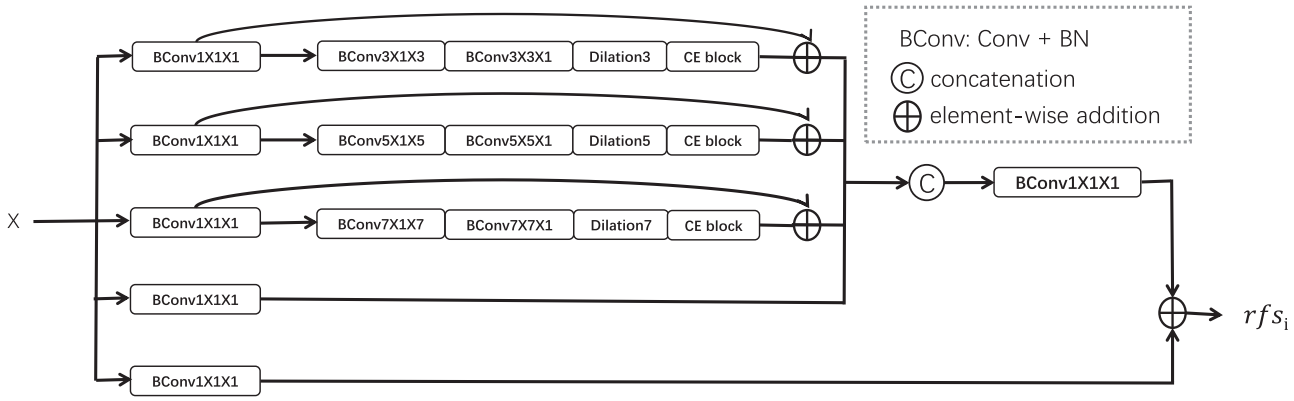

FIGURE A1 The flowchart of the Res-RF block. Res-RFs, Receptive Fields.

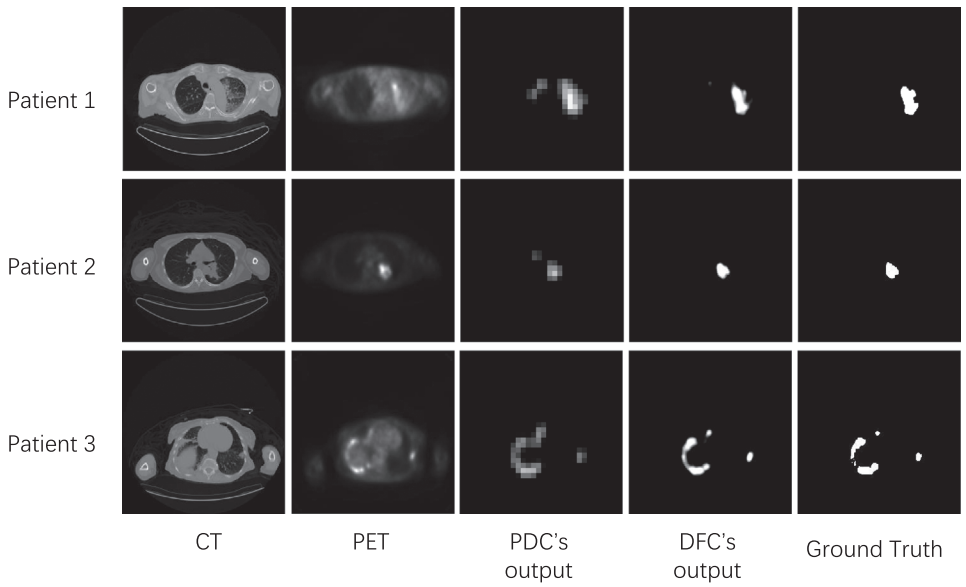

FIGURE A2 Illustrations of the outputs of the PDC and the DFC modules. DFC, decoding focusing component; PDC, partial decoder component.

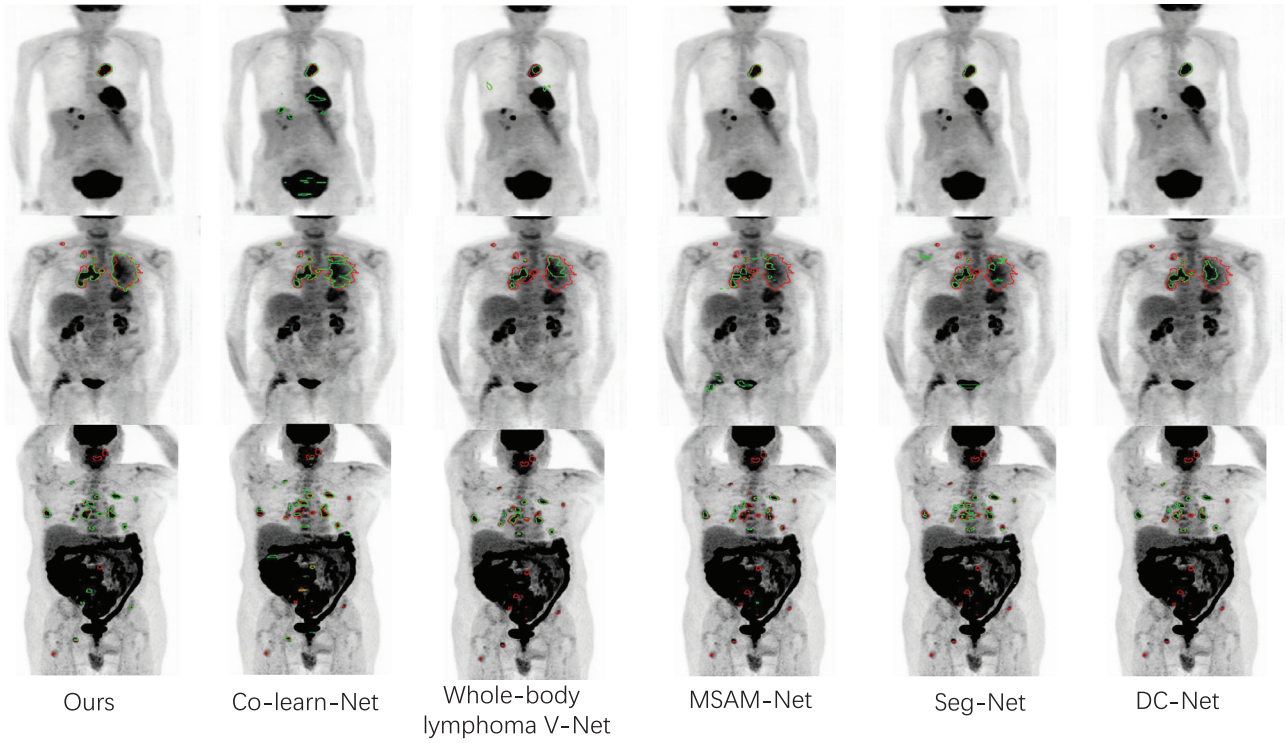

FIGURE A3 Three sample patients' MIPs from the front view with tumor contours delineated by radiologist (red) and different methods (green). MIPs, Maximum intensity projections.

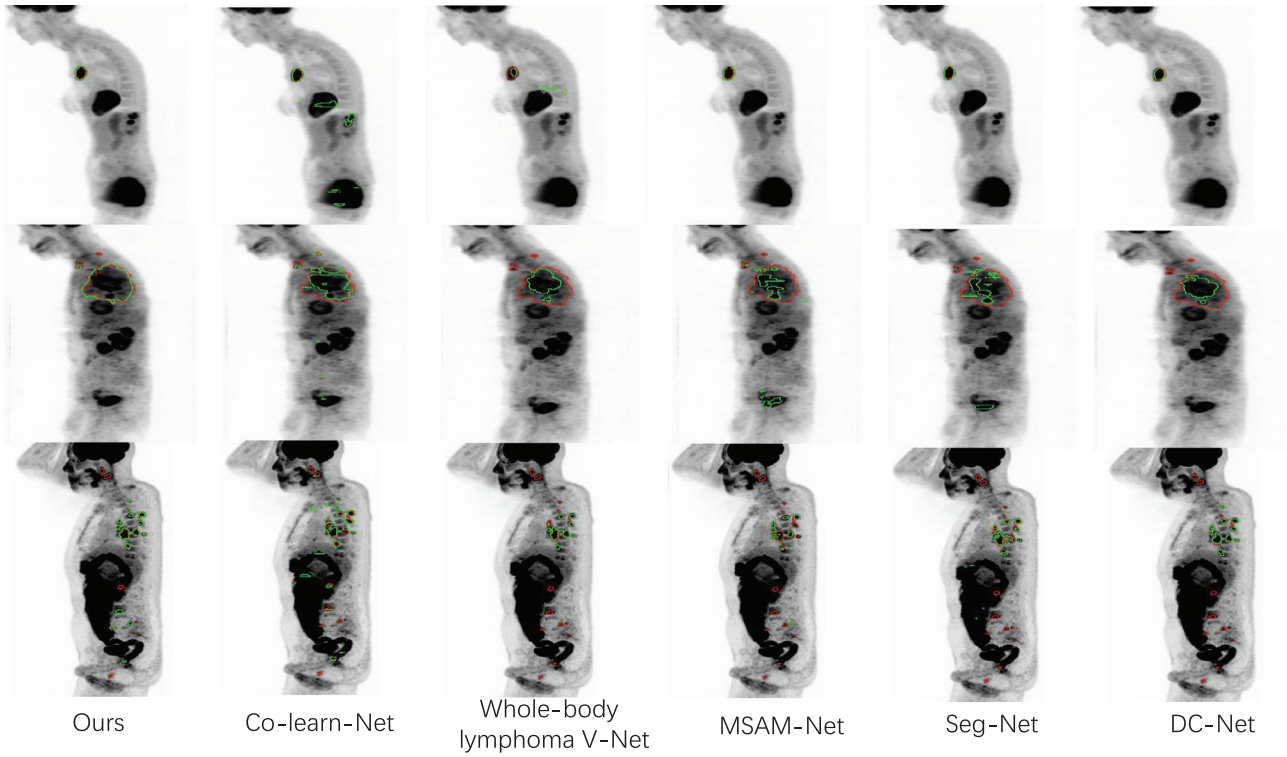

FIGURE A4 Three sample patients' MIPs from the side view with tumor contours delineated by radiologist (red) and different methods (green). MIPs, Maximum intensity projections..