IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 July 2022, accepted 24 July 2022, date of publication 1 August 2022, date of current version 8 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3195906

RESEARCH ARTICLE

Joint Lymphoma Lesion Segmentation and Prognosis Prediction From Baseline FDG-PET Images via Multitask Convolutional Neural Networks

PENG LIU[1], MIAO ZHANG[2], XIAORU GAO[1], (Graduate Student Member, IEEE), BIAO LI[2,3], AND GUOYAN ZHENG[1], (Member, IEEE)

[1]School of Biomedical Engineering, Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai 200240, China
[2]Department of Nuclear Medicine, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China
[3]Collaborative Innovation Center for Molecular Imaging of Precision Medicine, Ruijin Center, Shanghai 200025, China

Corresponding authors: Guoyan Zheng (guoyan.zheng@sjtu.edu.cn) and Biao Li (lb10363@rjh.com.cn)

This work was supported in part by the Natural Science Foundation of China under Grant U20A20199, in part by the Science and Technology Commission of Shanghai Municipality under Grant 20511105205, in part by the National Key Research and Development Program of China under Grant 2019YFC0120603, in part by the Medical and Engineering Cross Research Foundation of Shanghai Jiao Tong University under Project YG2019ZDA22 and Project YG2019ZDB09, and in part by the Shanghai Municipal Key Clinical Specialty under Project shslczdzk03403.

This work involved human subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Shanghai Ruijin Hospital Review Board and performed in line with the 1964 Helsinki Declaration.

ABSTRACT *Objective:* Lymphoma lesion segmentation and prognosis prediction from baseline FDG-PET images are valuable for tailoring and adapting a treatment plan for patients with Diffuse Large B-cell Lymphoma (DLBCL). However, the tasks are challenging due to the fact that DLBCL is a highly heterogeneous group of neoplasms and that the lymphoma cells are large and arranged in a diffuse pattern. *Methods:* We propose a novel multi-task 3D convolutional neural network model for simultaneous lymphoma lesion segmentation and prognosis prediction from baseline FDG-PET images. In our model, the learned image features of one task are shared and thereby mutually reinforce the learning of the other task. Since the dataset is limited, to reduce overfitting and to facilitate network convergence, we further introduce deep supervision into both the segmentation task and the prognosis prediction task. *Results:* Evaluated on a dataset of 269 patients, our method achieves an average Dice similarity coefficient of 0.868 for lesion segmentation, an average AUC (area under the curve) of 0.823 and an average accuracy of 0.821 for prognosis prediction. Its predictions can differentiate patients with different PFS (progression-free survival) and OS (overall survival) ($p < 0.0001$). *Conclusion:* Our method achieves joint lymphoma lesion segmentation and prognosis prediction from baseline FDG-PET scans. *Significance:* Our model may be used to asset the physician as a second opinion while making the final decision.

INDEX TERMS Diffuse large b-cell lymphoma (DLBCL), prognostic model, multi-task learning, PET images.

I. INTRODUCTION

Diffuse large B-cell lymphoma (DLBCL), which accounts for about 30% to 40% of non-Hodgkin lymphoma, is one of the most common subtypes of non-Hodgkin lymphoma and has obvious heterogeneity in morphology, clinical characteristics and prognosis [1]. The current treatment strategy for DLBCL is to use standard first-line treatment based on R-CHOP (rituximab, cyclophosphamide, doxorubicin, vincristine, and prednisone). Although most patients can be relieved after

The associate editor coordinating the review of this manuscript and approving it for publication was Cristian A. Linte.

treatment with this regimen, there are still 20% to 40% of patients who relapse, and most of the relapses occur within the first 2 years [2]. Early identification of patients with a poor prognosis allows for the tailing of their curative remediation plan for an improved chance of cure.

In DLBCL, pre-treatment prognosis is traditionally estimated using the International Prognostic Index (IPI) or one of its modifications [3], [4]. However, the role of these prognostic factors, which are based on tumor burden surrogates, is limited [4]. Positron Emission Tomography (PET) using 18F-fluorodeoxyglucose (18F-FDG) is a whole-body metabolic imaging technique that can accurately characterize tumor heterogeneity in a non-invasive manner and plays an important role in staging, treatment monitoring and prognostic evaluation of lymphoma. Among parameters measured from PET images, Standardized Uptake Value (SUV) is the most common for quantifying the degree of FDG uptake and glucose metabolism in sites of disease. Analysis and interpretation of 18F-FDG-PET images is challenging as it is usually performed by trained radiologists or readers who visually inspect the images slice by slice for tumors and then delineate multiple regions of interest (ROI) manually over each lymphoma lesion, in order to derive SUV-related quantitative measures. Manually based analysis can be very labor-intensive and time-consuming, especially in whole-body FDG-PET scans, and is operator dependent, suffering from intra- and inter-reader variability [5]. Moreover, interpretation of these quantitative measures is still controversial, reflected by the unstable manifestations of assessments and prognostic values reported in many studies [6]. Thus, effective computer-aided diagnosis (CAD) systems are essential in identifying high-risk patients who could benefit from intensive or novel therapies early.

Lymphoma lesion segmentation and prognosis prediction are two basic tasks in CAD systems. It has been previously shown that both high metabolic heterogeneity and total metabolic tumor volume (TMTV) have strong prognostic values before initiating therapy in DLBCL [7], [8]. Thus, shape properties are useful in both lesion segmentation and prognosis prediction. Consequently, it is worth to explore the training of two tasks jointly in one network to encourage feature sharing between tasks but this has never been done before for lymphoma lesion segmentation and prognosis prediction.

In this paper, we propose a novel multi-task 3D deep learning model to jointly train lesion segmentation and prognosis prediction from baseline FDG-PET Images. We use 2-year event-free survival (2y-EFS) as the gold standard for our prognosis prediction, as it is a robust end-point for disease-related outcomes in DLBCL treated with immunochemotherapy [2]. Since dataset is usually limited, to reduce overfitting and to facilitate network convergence, we further introduce deep supervision into both the segmentation task and the prognosis prediction task. The proposed method features an encoder-decoder network for lesion segmentation and a multi-scale classification network for prognosis prediction. We employ 3D U-Net as the backbone network for both lesion segmentation and prognosis prediction [9]. The lesion segmentation and the prognosis prediction tasks share features extracted from the encoding path. Our main contributions are summarized as follows:

1) We propose a novel multi-task deep learning model for simultaneous lymphoma lesion segmentation and prognosis prediction from baseline FDG-PET images. The proposed method uses an encoder-decoder network for lesion segmentation. Multi-scale features extracted from each level of the decoder path are fused for prognosis prediction.
2) We introduce deep supervision into both the segmentation task and prognosis prediction task. We show that deep supervision mechanism boosts not only the performance of prognosis prediction but also the performance of lymphoma lesion segmentation.
3) To the best knowledge of the authors, this is the first study which addresses joint lymphoma lesion segmentation and prognosis prediction from baseline FDG-PET images with an end-to-end CNN model and has a clear advantage over previous work where only single task is addressed.

In the following, we first summarize related work in Section II. We then present our methods in Section III. Data and experiments are presented in Section IV. We show our experimental results in Section V, followed by a discussion on Section VI. We conclude the paper in Section VII.

II. RELATED WORK

In this section, we review the relevant work on lymphoma lesion segmentation and prognosis prediction.

A. LYMPHOMA LESION SEGMENTATION

It is challenging to develop automatic methods for lymphoma detection and segmentation on whole-body FDG-PET images. Firstly, due to the distribution characteristics of the lymphatic system, shape and size of lymphomas vary greatly from location to location where different lymphomas have different SUVs. Especially in DLBCL, the high distribution variability of nodal and/or extranodal lesions among patients makes automatic detection and segmentation of lymphoma lesions an even more challenging task. Secondly, physiological FDG uptake (brain, myocardium, liver, brown fat, digestive tract) and radiopharmaceutical clearance (kidneys, urethras, bladder) cause that some normal organs have SUVs similar to target lesions. Conventional threshold-based methods can easily mis-classify these normal organs as lymphomas.

To meet the challenge, a large number of PET lesion segmentation methods have been developed, ranging from ROI-dependent methods to machine learning-based methods [10]–[15]. ROI-dependent methods require a physician to manually define a ROI, followed by segmentation using a thresholding by 40% or 50% of the maximum SUV (SUV_{max}) in ROI. This type of methods is time consuming due to the requirement of manual definition of ROI and generates poor

results as the SUV values in a ROI are not homogeneous. Different methods including clustering-based methods [13], conditional random fields [12] and level set methods [16] are proposed to improve it. Several studies have shown that automated algorithms such as conditional random fields or methods based on clustering or level set outperform thresholding methods [12], [13], [16]. Recently, machine learning-based methods have gained more and more interest. A crucial step in the design of such systems is the extraction of discriminant features from the images, which is usually done by human researchers [17]. The limited representation capability of these hand-crafted features makes it difficult to handle large variations of appearance and shape of distributed lymphoma lesions.

The more recent development of deep neural networks, and in particular convolutional neural networks (CNN) [18] suggests another course of method to solve the challenging lymphoma lesion segmentation tasks [19]. A recent study comparing 11 automated PET segmentation methods in lymphoma showed that 3D CNN, among other methods, achieved both good lesion-level segmentation and patient-level quantification performance [20]. Contrary to conventional shallow learning methods, where feature design is crucial, deep learning methods automatically learn hierarchies of relevant features directly from the training data. Based on an industrial software prototype, Capobianco *et al.* [21] presented a study for automatic delineation and classification of lymphoma lesions on PET/CT images. Jemaa *et al.* [5] proposed to use cascaded 2D and 3D CNN for automatic tumor segmentation and feature extraction from whole-body FDG-PET/CT images. Blanc-Durand *et al.* [22] investigated the performance of the open source 3D CNN architecture called nnU-Net [23] on segmenting TMTV in large datasets of patients with DLBCL. Xu *et al.* [24] proposed a deep dilated convolutional encoder-decoder architecture for lymph node segmentation on PET/CT images while Wang *et al.* [25] proposed to leverage spatial-temporal correlation between the decorder feature maps for lymphoma segmentation. Although these methods have been successfully applied to automatic detection and delineation of lymphoma lesion from PET images [5], [20]–[22], [24], [25], they suffer from the limitation that these methods only do single task and do not exploit useful information contained in multiple related tasks such as segmentation and classification. Thus, the performance of these methods can be further improved.

B. PROGNOSIS PREDICTION FROM BASELINE FDG-PET IMAGES

Accurate pretreatment evaluation and response assessment are critical to the optimal management of patients with DLBCL. Much efforts have been made to identify accurate prognostic imaging biomarkers or radiomics features extracted from baseline FDG-PET scans. For example, Sasanelli *et al.* [7] calculated TMTV on baseline 18F-FDG-PET/CT scans of 114 patients diagnosed with DLBCL and used Cox proportional hazard regression model to show that TMTV was an independent prognostic predictor. Nguyen *et al.* [26] found that the SUV_{max} and the maximum tumor diameter parameters of 18F-FDG PET images were useful indicators of DLBCL prognosis. Zhang *et al.* [6] combined the parameters of baseline and interim 18F-FDG-PET/CT scans of 85 DLBCL patients, demonstrating the independent prognostic abilities of baseline Tumor Lesion Glycolysis (TLG) and $\triangle$ SUV_{max}. Aide *et al.* [27] verified that LZHGE (Long-Zone High-Grey Level Emphasis) texture features extracted from baseline 18F-FDG PET images were independent predictors of 2y-EFS of DLBCL. Zhang *et al.* [28] developed an analytic approach combining radiomics signature and International Prognostic Index (IPI) to predict the progression-free survival and overall survival of patients with DLBCL. Eertink *et al.* [29] investigated the added value of baseline radiomics features to the IPI in predicting outcome after the first-line treatment. Although encouraging results are reported, most of these methods may be too simple, assuming that the outcome is a linear combination of covariates. Thus, the prediction accuracy of these methods may suffer from the remarkable complexity of DLBCL.

Recently, more and more studies reported that deep CNNs were superior to standard survival analysis [21], [30]. Compared with manually drafted features, deep learning features are learned directly from data, which are more adapted to specific tasks and more naturally correlate with clinical results [31]. Each convolutional layer of deep learning networks contains hundreds of convolutional filters, which can describe multi-level tumor information ranging from low-level visual features to high-level abstract features. For lymphoma, there are studies using deep learning to classify 18F-FDG-PET/CT uptake patterns [21], [30]. But to the best knowledge of the authors, there is no study applying deep learning methods on baseline FDG-PET images for predicting 2y-EFS in patients with DLBCL, enabling outcome prognostication directly from baseline FDG-PET scans.

III. METHODS

In this paper, we propose a multi-task learning framework. The proposed framework integrates lymphoma lesion segmentation and prognosis prediction in an end-to-end CNN model which takes FDG-PET volumes as inputs and generates two outputs for each feature level, including a segmentation map and a probability of 2y-EFS.

A. MULTITASK LEARNING CNN

The proposed CNN is shown in Fig. 1. We adopt 3D U-Net [9] as the backbone network as it achieves excellent performance in 3D medical image analysis. There are three parts in the 3D U-Net architecture: 1) an encoding path, 2) a decoding path, and 3) skip connections between them. As shown in Fig.1, the encoder path utilizes three down-sampling operations and focuses on analysis and feature representation learning from the input data. Symmetrically, there are three up-sampling operations in the decoding path, which generates

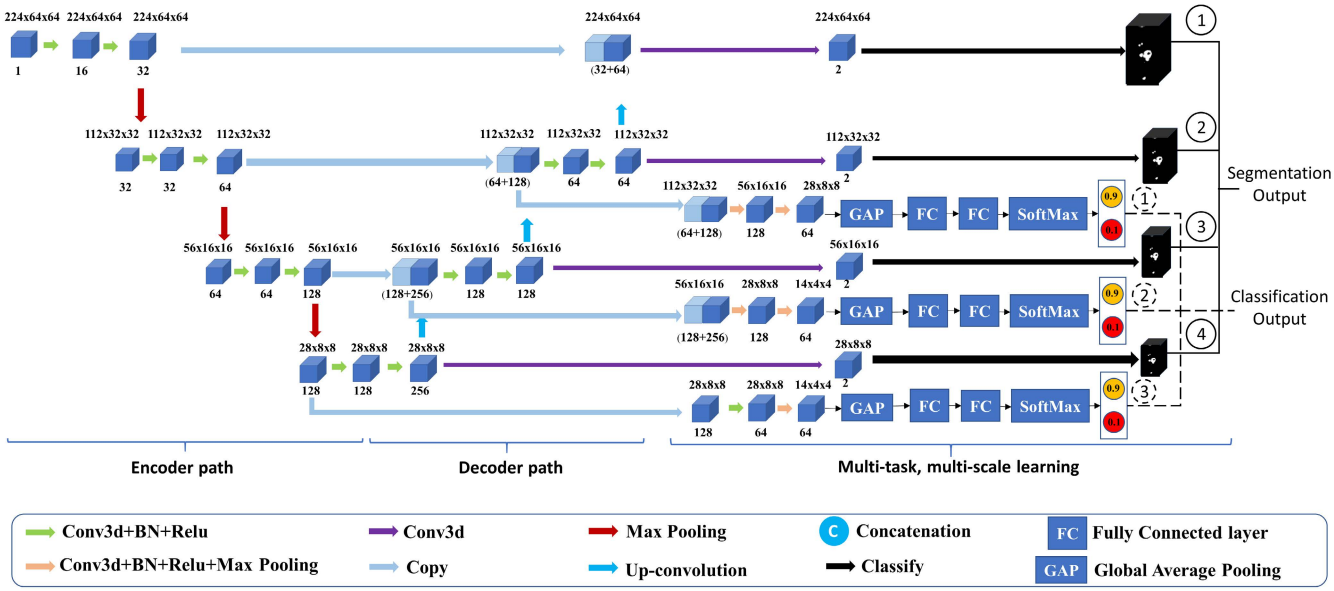


FIGURE 1. A schematic illustration of the overall architecture of our method, which is a multi-task U-Net with deep supervision on both tasks. The number above each block shows the data dimension while the number below each block represents the number of feature maps.

segmentation prediction, relying on the learned features from the encoder path. Shortcut connections are established between layers of equal resolution in the encoder and decoder paths to propagate spatial information and to facilitate forward and backward information flow. In the proposed CNN, all convolutional layers use a kernel size of $3 \times 3 \times 3$ and stride of 1 and all max pooling layers uses a kernel size of $2 \times 2 \times 2$ and stride of 2. In the convolutional and deconvolutional blocks of our network, batch normalization (BN) and rectified linear unit (ReLU) are adopted to speed up the training and to enhance the gradient back-propagation.

High-level abstract features learned by CNNs are used in many classification algorithms [18], [32]–[34]. Furthermore, multi-scale features can be used to incorporate both local and larger contextual information for an improved performance [35], [36]. Inspired by these observations, we propose to share high-level feature maps from U-Net for both lymphoma lesion segmentation and prognosis prediction in the multi-task learning framework. We add a classification branch for each resolution of the decoder path of the 3D U-Net as shown in Fig. 1 to predict prognosis. Specifically, high-level features extracted from each resolution along the decoder path are fed into the classification branch. Each classification head consists of two fully connected (FC) layers and one softmax layer to predict the 2y-EFS probability. Then we fuse these predictions of the classification branch of different resolutions for the final prognosis prediction task.

B. MULTILEVEL DEEP SUPERVISION

In the proposed multi-task learning framework, we share the encoder path for prognosis prediction and lymphoma lesion segmentation tasks to extract common features for these two tasks. Because of the multi-scale characteristics of lymphoma lesions, we design a deep supervision architecture for both the segmentation task and the classification task as illustrated in Fig. 1. Deep supervision is helpful to reduce overfitting and facilitate network convergence when training a deep neural network. It is also useful to extract more meaningful features. Here, deep supervision is utilized in each stage, so that the output of the middle stage can be directly utilized as a supervision. For segmentation task, the output of each decoder stage is followed with a convolutional layer, which uses a kernel size of $3 \times 3 \times 3$ and stride of 1. After the convolutional layer, we can get a 2-channel feature map, followed by applying softmax operation to get the segmentation result of this stage. We then compute segmentation loss use the segmentation result and the down-sampled ground truth segmentation mask as shown in Fig. 1. For classification task, except the feature maps at the lowest and the highest resolutions of the 3D U-Net, all other feature maps extracted from each resolution along the decoder path are further processed with two convolutional operations followed with BN, ReLU and max pooling to reduce the spatial resolution. For the feature maps at the lowest resolution of the 3D U-Net, we only applied once max pooling while all other operations are the same as other feature maps. After that, we propose to use spatial-wise global average pooling (GAP) to convert feature maps from different resolutions to the same size. GAP layer transforms feature maps of size $(D \times H \times W \times C)$ to feature maps of size $(1 \times C)$ by simply taking the average of each channel where $(D \times H \times W)$ is the size of the volume, and C is the number of channels. Although we can also use global max pooling (GMP) here, empirically we find that GAP leads to better results than GMP, probably due to the fact that GMP only uses the max value to represent the whole feature map, disregarding useful spatial information.

To this end, we obtain a 64-dimensional feature vector from the feature maps of each resolution. These feature vectors are then fed into the classification head of each decoder stage to get the prognosis prediction. As shown in Fig. 1, we can get three prognosis predictions in total and we will combine those predictions with a weighted average strategy to get the final prognosis prediction.

C. MULTITASK AND DEEP SUPERVISION LOSS FUNCTION

Class imbalance is common in medical image classification tasks. For example, in our dataset, the number of cases with 2-year complete remission (2y-CR) is more than twice that of 2y-EFS. To address the issue, we adopt a weighted cross entroy loss as follows:

$$L_{cls}(p_{cls}, y_{cls}) = -w_p y_{cls} log(p_{cls}) - w_n(1-y_{cls}) log(1-p_{cls}) \tag{1}$$

$$p_{cls} = w^1_{cls} p^1_{cls} + w^2_{cls} p^2_{cls} + w^3_{cls} p^3_{cls} \tag{2}$$

where $p^i_{cls}, i = 1, 2, 3$ is the predicted 2y-EFS probability from each stage of decoder path. $w^i_{cls}, i = 1, 2, 3$ is the weight of each stage. We empirically set w^1_{cls}, w^2_{cls} and w^3_{cls} as 0.1, 0.2, 0.7, respectively. y_{cls} is the ground-truth label of this volume ($y_{cls} = 0$ for 2y-CR and $y_{cls} = 1$ for 2y-EFS). w_p and w_n are weights for 2y-EFS cases and for 2y-CR cases, respectively. They are defined as:

$$w_p = \frac{N_n}{N_p + N_n}, \quad w_n = \frac{N_p}{N_p + N_n} \tag{3}$$

where N_p and N_n are the numbers of 2y-EFS cases and 2y-CR cases, respectively.

For the segmentation task, there exists imbalance between foreground and background, which may cause segmentation bias. To account for this, we use a segmentation loss based on the Dice coefficient between the predicted segmentation maps and ground truth segmentation masks. For stage i, it is defined as:

$$L^i_{seg}(p^i_{seg}, y^i_{seg}) = 1 - \frac{2p^i_{seg} y^i_{seg} + 1}{p^i_{seg} + y^i_{seg} + 1} \tag{4}$$

where p^i_{seg} and y^i_{seg} denote the predicted segmentation maps from the proposed CNN model and the ground truth segmentation masks of stage i, $i = 1, 2, 3, 4$, respectively. Therefore, the overall segmentation loss of our deep supervision model is the weighted sum of segmentation loss of each stage:

$$L_{seg} = w^1_{seg} L^1_{seg} + w^2_{seg} L^2_{seg} + w^3_{seg} L^3_{seg} + w^4_{seg} L^4_{seg} \tag{5}$$

where $w^i_{seg}, i = 1, 2, 3, 4$ is the loss weight of each stage. We empirically set w^1_{seg}, w^2_{seg}, w^3_{seg} and w^4_{cls} as 0.53, 0.27, 0.13, 0.07, respectively.

Our multi-task learning loss is then a linear combination of the classification loss L_{cls} and the segmentation loss L_{seg} by a hyperparameter λ:

$$L_{mul} = \lambda L_{seg} + (1-\lambda) L_{cls} \tag{6}$$

where L_{mul} is our multi-task learning loss, and λ is the weight of the segmentation task. We empirically set $\lambda = 0.1$.

IV. DATA AND EXPERIMENTS

All study procedures were approved by the Ethics Committee of Shanghai Ruijin Hospital, Shanghai Jiao Tong University School of Medicine. All procedures performed in this study involving human participants were in accordance with the ethical standards of Ruijin Hospital, Shanghai Jiao Tong University School of Medicine research committee and with the Helsinki declaration and its later amendments or comparable ethical standards.

A. DATA DESCRIPTION

1) PATIENT POPULATION

We conducted a retrospective review of patients with DLBCL obtained from the database of the department of nuclear medicine of Shanghai Ruijin Hospital for the period from January 2013 to February 2018. All the patients undertook baseline 18F-FDG-PET/CT scans. The inclusion criteria were as follows: 1) New diagnosis of lymphoma confirmed by histology; 2) Availability of conventional imaging and non-imaging data for staging disease; 3) 18F-FDG-PET/CT scan performed at staging, with evidence of FDG-avid disease; 4) All patients were treated by first-line immunochemotherapy; 5) Availability of clinical information and/or imaging information for validation of patient PET findings; 6) Post-therapy surveillance, until death or for at least 24 months after the diagnosis to evaluate the risk of progression and the survival rate. In total we have 269 patients with 2y-EFS labels (89 cases with 2y-EFS and 180 cases with 2y-CR).

2) PET IMAGING AND SEGMENTATION PROTOCOL

All 18F-FDG-PET/CT images were acquired using a GE Discovery VCT PET/CT scanner. The radiochemical purity of 18F-FDG was over 95% (provided by Shanghai Atomic Kexing Pharmaceutical Co., Ltd). The patients were fasted for at least 6 hours before the PET study. Serum glucose level was determined at the time of FDG injection using a glucometer, and all patients demonstrated a glucose level below 7.8mmol/L. PET scan was performed 60 min after intravenous administration of 0.15mCi/kg of FDG. CT acquisition was performed with the following parameters: 60 mAs, 140 kVp, pitch 1. Subsequently, PET images were acquired from the base of the skull to the mid-thighs (legs were included when necessary). All examinations were reconstructed using the Ordered Subset Expectation Maximization (OSEM) algorithm with point spread function (PSF) modelling with three iterations and 21 subsets with filtering. Scatter and attenuation corrections were applied.

After image acquisition, lymphoma lesions were manually delineated from the acquired PET/CT data by four experienced physicians using open-source software ITK-SNAP 3.8 (ITK-SNAP 3.x, http://www.itksnap.org/). All the physicians have been certified by two boards (nuclear medicine and radiology) including two senior physicians having more than 20 years' working experience. Majority voting was used to

TABLE 1. Data distribution in each fold.

Data	Fold#1	Fold#2	Fold#3	Fold#4	Fold#5
2y-EFS	18	18	17	17	19
2y-CR	35	36	37	37	35

assign voxel labels. In total physicians manually delineated lymphoma lesions of 269 patients.

B. STUDY DESIGN

1) MAIN STUDY

We first conducted a five-fold cross-validation study to compare the proposed method with other state-of-the-art (SOTA) methods. The data distribution of the five folds is shown in Table 1. Each time, data in one fold were used as the testing data while the data in the remaining four folds were used for training. As there was no previous work on joint segmentation and prognosis prediction from baseline FDG-PET images, we compared our method with a simple multi-task U-Net without deep supervision as shown in Fig. 2-(b) and the multi-task V-Net method as introduced in [37], which generalized V-Net for joint segmentation and classification of tumors in 3D breast ultrasound images. It fused the feature maps from the encoder layer before the bottom of the V-Net, the feature maps at the bottom of the V-Net and the feature maps from the decoder layer right after the bottom of the V-Net and then fed the fused feature maps to a second branch to generate the classification label. This method has been shown in [37] to generate better segmentation and classification results than other SOTA multi-task learning methods [38], [39], when applied to 3D breast ultrasound images.

We compared the segmentation performance of our proposed method with 3D U-Net, which was regarded as a baseline method for medical image segmentation. We additionally compared the classification performance of our proposed method with ResNet34 [33] and DenseNet121 [40], which were the SOTA deep learning methods for classification.

2) ABLATION STUDY

In this study, we took the data in a randomly selected fold as testing data and all the data in the remaining folds as training data. We conducted ablation study to investigate the effectiveness of individual components of our proposed method. Specifically, we conducted following ablation experiments: 1) single-task vs. multi-task; 2) multi-task without deep supervision vs. multi-task with deep supervision; 3) multi-task with deep supervision on a single task vs. multi-task with deep supervision on both tasks (Ours). Table 2 summarizes different CNN models that we investigated. We treated the segmentation U-Net (model-A) and the classification U-Net-Encoder (model-B) as shown in Fig. 2-(a) as single-task models. For the classification U-Net-Encoder, we only used the encoder path from the U-Net, followed by a GAP layer and a classification head consisting of two FC layers and one softmax layer to predict the 2y-EFS probability. For

TABLE 2. Different CNN models used in our ablation study. SEG: segmentation; CLS: classification.; DP: Deep supervision; w/o: without; w/: with.

CNN Models	Tasks	DP on CLS	DP on SEG
SEG U-Net (model-A)	SEG	-	No
CLS U-Net-Encoder (model-B)	CLS	No	-
Multi-task U-Net w/o DP (model-C)	SEG + CLS	No	No
Multi-task U-Net w/ DP on SEG (model-D)	SEG + CLS	No	Yes
Multi-task U-Net w/ DP on CLS (model-E)	SEG + CLS	Yes	No
Ours	SEG + CLS	Yes	Yes

multi-task learning, we first investigated a multi-task U-Net without deep supervision (model-C) as shown in Fig. 2-(b), which generalized 3D U-Net to joint segmentation and classification. This was done by feeding the feature maps at the bottom of the 3D U-Net to a second branch to predict the classification label. We further investigated two different multi-task CNN models with deep supervision. The first one (model-D) as shown in Fig. 2-(c) was designed to have deep supervision only on segmentation task while the second one (model-E) as shown in Fig. 2-(d) was designed to have deep supervision only on classification task. We finally compared the performance of all above mentioned CNN models with the proposed method.

3) MODEL TRAINING AND TESTING

The size of all FDG-PET images used in our study is $257 \times 128 \times 128$ voxels with a voxel spacing of $3.3mm \times 5.5mm \times 5.5mm$. We first cropped all images to a size of $247 \times 80 \times 64$ voxels in order to remove the region with empty area. After cropping, we then resized all images to $224 \times 64 \times 64$ voxels for training and testing. We trained all networks with a batch size 12. The learning rate was originally set as $1e-4$ and decayed with a power of 0.1 after 100 epochs. In total we trained all networks 200 epochs. Data augmentation was used to enlarge the training samples by rotating each image around three axes with a random angle sampled from the range from $-8°$ to $8°$ and by scaling each image along three axes with a random coefficient sampled from the range of (0.8∼1.2).

All models were implemented using Python with PyTorch library. We trained and tested all models on 4 NVIDIA tesla V100 GPUs. Adam optimizer [41] was used to train all models.

C. METRICS

For lesion segmentation, assuming the automatically segmented set of lesion voxels as AS and the manually defined ground truth lesion as GT, we employs Dice similarity coefficient (DSC), precision and recall to evaluate the performance of different segmentation models. These metrics are calculated as follows:

- *Dice Similarity Coefficient (DSC)* - it quantifies the match of two sets by normalizing the size of their intersection over the average of their sizes and is defined as follows:

$$DSC = \frac{2|AS \cap GT|}{|AS| + |GT|} \tag{7}$$

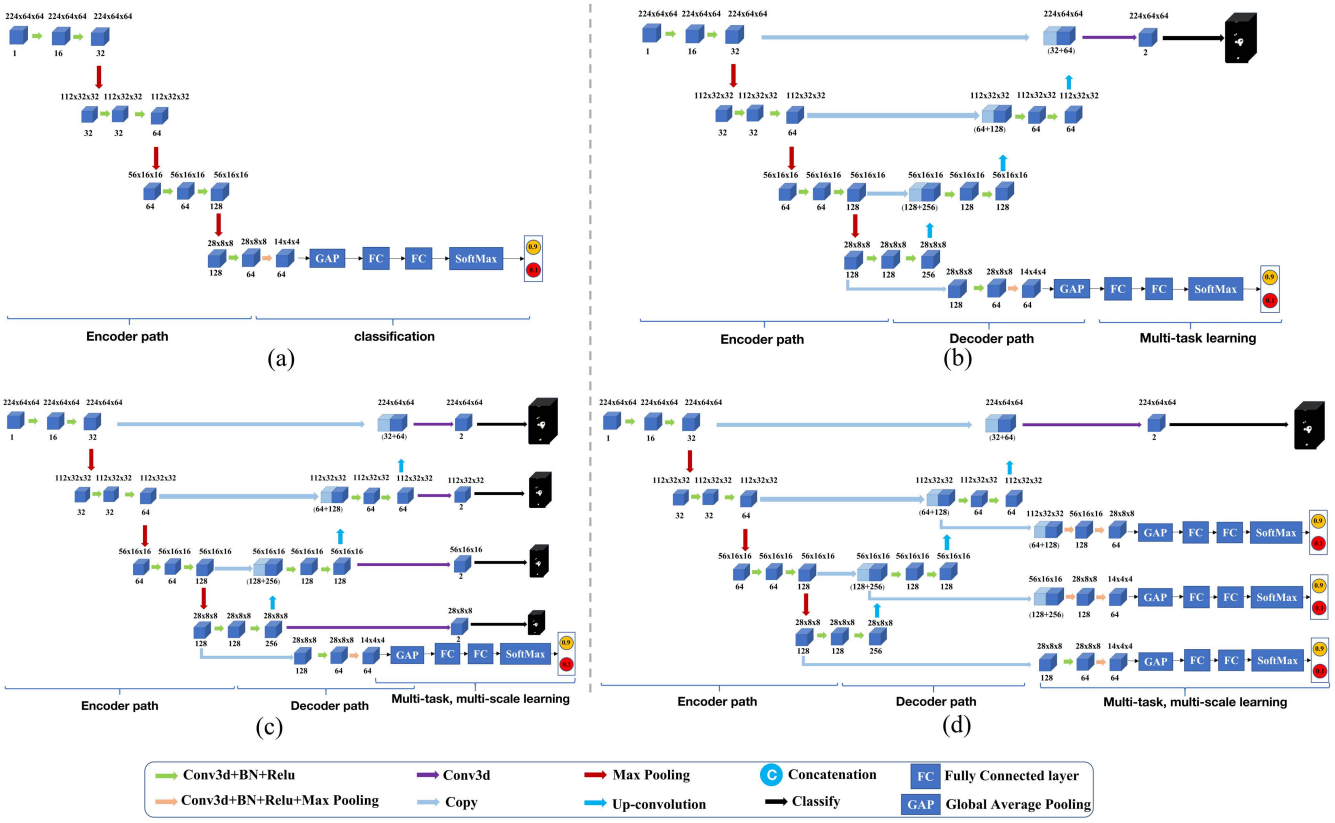


FIGURE 2. Illustration of different CNN models used in our ablation study. (a) U-Net-encoder for classification (model-B), (b) Multi-task U-Net without deep supervision (model-C), (c) Multi-task U-Net with deep supervision on segmentation (model-D), (d) Multi-task U-Net with deep supervision on classification (model-E).

where the operator $|\cdot|$ returns the number of voxels contained in a region.

- *Precision (PR)* - It is defined as the fraction of all automatically segmented lesion voxels that are correct:

$$PR = \frac{|AS \cap GT|}{|AS|} \tag{8}$$

- *Recall (RC)* - It is defined as the fraction of all ground truth lesion voxels that have been corrected segmented by an automatic method:

$$RC = \frac{|AS \cap GT|}{|GT|} \tag{9}$$

For prognosis prediction, performance metrics include receiver operating characteristic (ROC), area under ROC curve (AUC), sensitivity (SEN), specificity (SPE), and accuracy (ACC). Accuracy is a measure of the error rate (ratio of correct predictions to all predictions made). The ROC curve describes the true-positive rate (sensitivity) versus the false-positive rate (100% - specificity) at various thresholds, and an AUC of 100% represents a perfect test while an AUC of 50% indicated random predictions. Differences in sensitivity and specificity between different models are compared using the Youden test.

Statistical significance is considered at $p < 0.05$. Data are analyzed using SPSS software (IBM SPSS Statistics for Windows, version 25; Amonk, NY; IBM Corp.).

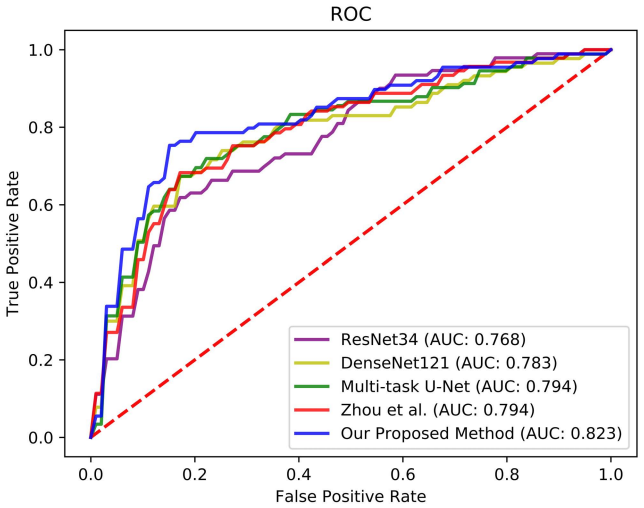


FIGURE 3. ROC and AUC of different methods when evaluated with 5-fold cross-validation study.

V. RESULTS

A. RESULTS OF MAIN STUDY

Table 3 shows the performance comparison of different methods in the main study. For lesion segmentation, the average DSC, prediction, and recall achieved by the proposed method are 0.868, 0.875, and 0.875, respectively. Our method outperforms the baseline method 3D U-Net and the Multi-task

TABLE 3. Main study results. SEG: segmentation; CLS: classification: deep supervision; w/o: without.

Method	Segmentation results			prognosis prediction results			
	DSC	Precision	Recall	Accuracy	Sensitivity	Specificity	Youden's index
CLS ResNet34 [33]	-	-	-	0.769	0.584	0.861	0.445
CLS DenseNet121 [40]	-	-	-	0.773	0.585	**0.867**	0.452
SEG U-Net [9]	0.844	0.884	0.849	-	-	-	-
Multi-task U-Net w/o DP	0.855	**0.900**	0.852	0.784	0.639	0.855	0.495
Zhou et al. [37]	0.853	0.891	0.853	0.792	0.651	0.861	0.512
Our method	**0.868**	0.875	**0.875**	**0.821**	**0.730**	0.866	**0.596**

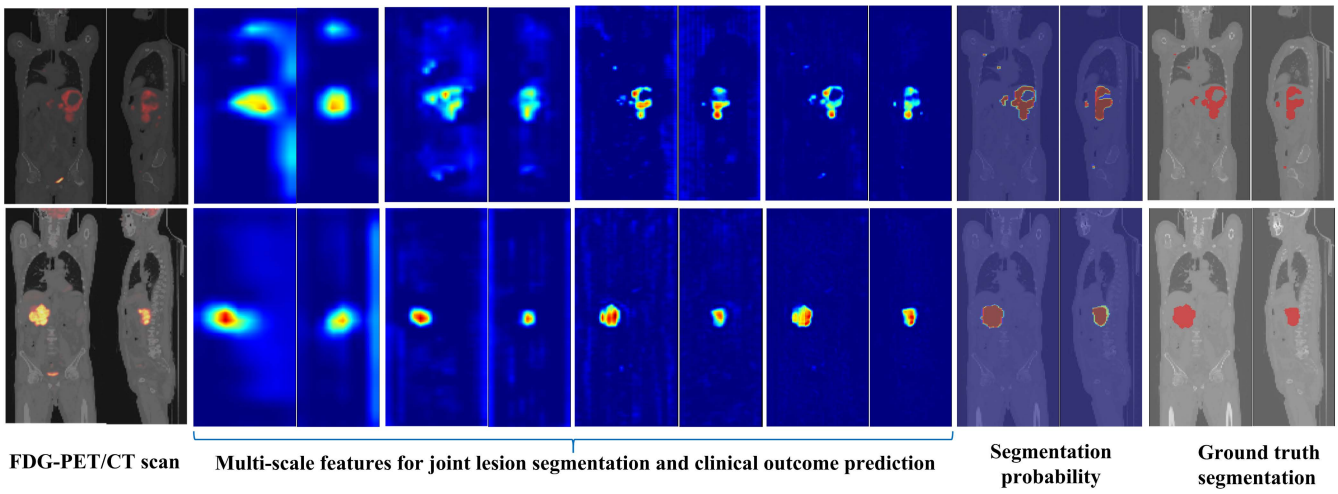


FIGURE 4. Coronal and sagittal views of multi-scale features for joint lesion segmentation and prognosis prediction. Left column: the input FDG-PET/CT scans; middle four columns: multi-scale features extracted from different resolutions; the fixth column: segmentation probability obtained by our method; the right column: ground truth segmentation. The ground truth label for the case shown in the top row is 2y-EFS and our system correctly predicted the case as 2y-EFS with a probability of 0.801. We also obtained a DSC of 0.917 for lesion segmentation. Similarly, the ground truth label for the case shown in the bottom row is 2y-CR and our system correctly predicted the case as 2y-EFS with a probability of 0.315 (i.e., predicted as 2y-CR with a probability of 0.685). A DSC of 0.922 was achieved by our method for lesion segmentation.

U-Net (without deep supervision). In contrast, the SOTA multi-task V-Net method introduced in [37] achieved an average DSC of 0.853, an average precision of 0.891 and an average recall of 0.853. Except average precision, our method achieved better segmentation results in terms of other two metrics than the method introduced in [37]. Fig.5 shows the 3D renderings of prediction segmentation of some typical cases. Green rectangle shows the difference between those 3D renderings. We can see that our method get better segmentation result when compared with other methods.

In regards to prognosis prediction, our method achieved much better results than the ResNet34 [33], the DenseNet121 [40], the Multi-task U-Net (withou deep supervision) and the SOTA multi-task V-Net method [37]. Specifically, our method achieved an average accuracy of 0.821, an average sensitivity of 0.730, an average specificity of 0.861 and a Youden's index of 0.596. In contrast, the method introduced in [37] achieved an average accuracy of 0.792, an average sensitivity of 0.651, an average specificity of 0.861 and a Youden's index of 0.512 while the Multi-task U-Net (without deep supervision) achieved worse results with an average accuracy of 0.784, an average sensitivity of 0.639, an average specificity of 0.855 and a Youden's index of 0.495. Our method also achieved better results than ResNet34 [33] and DenseNet121 [40]. To identify the prognostic value of different models, we draw ROCs of all methods to compare the power between the proposed method and other two SOTA methods, as shown in Fig. 3. The AUC of the proposed method was 0.821, which was better than other SOTA methods.

To further investigate the learning behavior of the proposed method, we draw the coronal and sagittal views of multi-scale feature maps from different resolutions. Fig. 4 displays the joint lesion segmentation and prognosis prediction results for two cases (Case 1: ground truth label as 2y-EFS and Case 2: ground truth label as 2y-CR) together with the multi-scale features. The proposed method can suppress the responses from physiological uptake and enhance the responses from lesion regions, leading to accurate lesion segmentation and prognosis prediction.

The prognostic value of the proposed method was further demonstrated by drawing the Kaplan-Meier curves of PFS (progression-free survival) and OS (overall survival) according to the predictions obtained by the proposed method. As shown in Fig. 6, the predictions of the proposed method were able to differentiate patients with different PFS and OS ($p < 0.0001$).

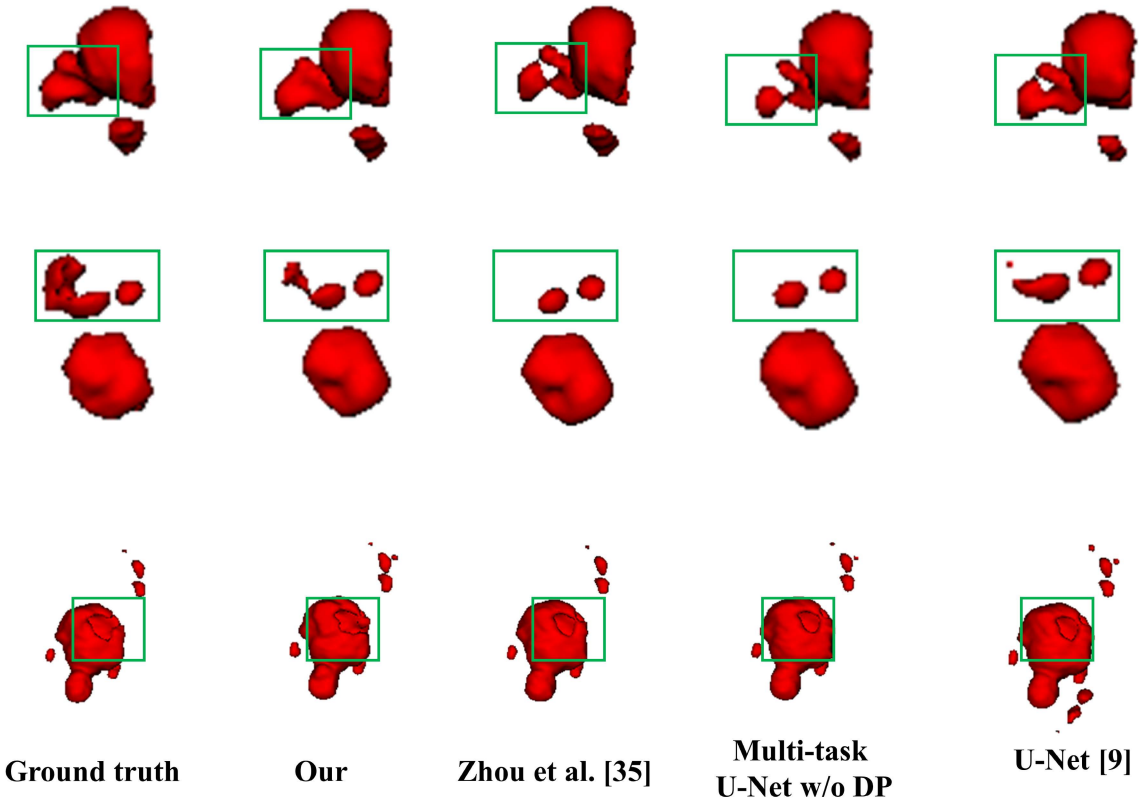


FIGURE 5. 3D renderings of segmentation results. Green rectangle shows the difference between those 3D renderings.

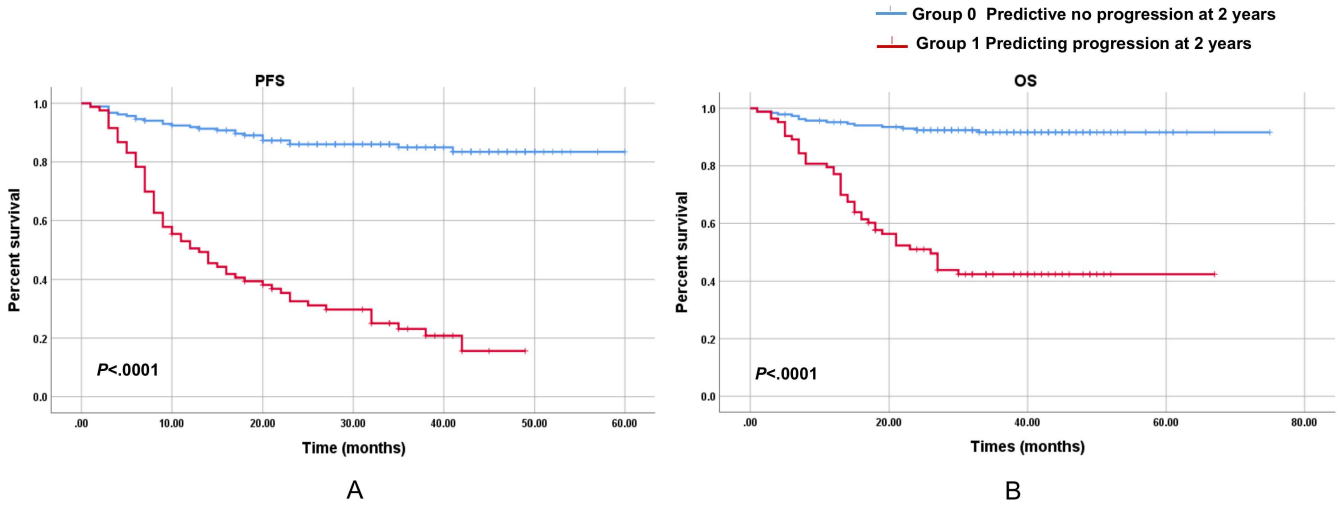


FIGURE 6. Kaplan-Meier estimates of progression-free survival (PFS) (A) and overall survival (OS) (B) of patients with DLBCL grouped by the proposed method's predictions.

B. RESULTS OF THE ABLATION STUDY

Table 4 presents the ablation study results. For the two baseline models, i.e., the 3D U-Net (model-A) [9] for lesion segmentation and the U-Net-Encoder (model-B) for prognosis prediction, the average DSC, precision and recall achieved by the 3D U-Net are 0.847, 0.864, and 0.864, respectively, while the average AUC, accuracy, sensitivity, specificity, and Youden's index achieved by the U-Net-Encoder are 0.784, 0.777, 0.588, 0.865 and 0.453, respectively. For model-C, which is a multi-task U-Net without deep supervision, the average DSC, precision and recall for lesion segmentation is 0.857, 0.878, 0.870, respectively, and the average AUC, accuracy, sensitivity, specificity, and Youden's index for classification is 0.804, 0.796, 0.647, 0.865, 0.512, respectively.

TABLE 4. Ablation study results. SEG: segmentation; CLS: classificaiton; DP: deep supervision; w/o: without; w/: with.

Method	Segmentation results			prognosis prediction results				
	DSC	Precision	Recall	AUC	Accuracy	Sensitivity	Specificity	Youden's index
SEG U-Net (model-A)	0.847	0.864	0.864	-	-	-	-	-
CLS U-Net encoder (model-B)	-	-	-	0.784	0.777	0.588	0.865	0.453
Multi-task U-Net w/o DP (model-C)	0.857	0.878	0.870	0.804	0.796	0.647	0.865	0.512
Multi-task U-Net w/ DP on SEG (model-D)	0.872	0.892	0.868	0.800	0.796	0.588	0.892	0.480
Multi-task U-Net w/ DP on CLS (model-E)	0.860	0.883	0.865	**0.831**	**0.833**	0.706	0.892	0.598
Our method	**0.878**	**0.896**	**0.879**	0.827	**0.833**	**0.706**	**0.892**	**0.598**

TABLE 5. Comparison of our method with SOTA methods in lymphoma segmentation and outcome prediction. SEG: segmentation; CLS: classification; MT: multi-task; DP: deep supervision.

Source	Task	Number of cases	Method	Performance	
				SEG (DSC, %)	CLS (AUC, %)
Blanc-Durand et al. [22]	SEG	733	3D CNN	0.730	-
Xu et al. [24]	SEG	63	2D U-Net	0.770	-
Wang et al. [25]	SEG	102	3D U-Net	0.859	-
Eertink et al. [29]	CLS	317	Radiomics	-	0.790
Zhang et al. [28]	CLS	152	Radiomics	-	0.789
Ours	SEG+CLS	269	3D MT-U-Net with DP	0.868	0.823

The results demonstrated that incorporating multi-task learning was helpful for both lesion segmentation and prognosis prediction. Model-D combines multi-task learning with deep supervision on segmentation, which achieved an average DSC of 0.872, an average precision of 0.892 and an average recall of 0.868 for lesion segmentation, and an average AUC of 0.800, an average accuracy of 0.796, an average sensitivity of 0.588, an average specificity of 0.892, and a Youden's index of 0.480 for prognosis prediction. The results demonstrated that incorporating deep supervision on segmentation task was helpful for lesion segmentation. In contrast, model-E combines multi-task learning with deep supervision on classificaiton, which achieved an average DSC of 0.860, an average precision of 0.883 and an average recall of 0.865 for lesion segmentation, and an average AUC of 0.831, an average accuracy of 0.833, an average sensitivity of 0.706, an average specificity of 0.892, and a Youden's index of 0.598 for prognosis prediction. The results demonstrated that incorporating deep supervision on classification task was helpful for prognosis prediction. Our method combines multi-task learning with deep supervision on both segmentation and classification tasks, achieving the best result on lesion segmentation and an equivalent performance to model-E on prognosis prediction.

VI. DISCUSSION

The treatment of patients with DLBCL is expensive. Moreover, compared with patients without relapse, the medical resource utilization and the cost of patients who have progressed/relapsed after first-line treatment are higher [3]. The high degree of heterogeneity of DLBCL poses unique challenges for predicting its prognosis. PET is a non-invasive imaging modality that can provide both functional and metabolic information of lymphoma lesions and is recommended for staging, restaging, therapy response assessment and recurrence detection of cancers [13]. However, quantitative analysis and interpretation of FDG-PET images is challenging as it usually requires labor-intensive and time-consuming manual delineation of multiple lymphoma lesions. Thus, it is essential to design CAD systems for FDG-PET images. Because of the large lesion shape and size variations, the high distribution variability of nodal and/or extranodal lesions, the existence of high SUVs in normal organs caused by the physiological FDG uptake and radiopharmaceutical clearance, lymphoma lesion segmentation and prognosis prediction are two challenging tasks.

In this paper, we proposed a multi-task 3D CNN mdoel for simultaneous lymphoma lesion segmentation and prognosis prediction from baseline FDG-PET images. We additionally introduced deep supervision on both lesion segmentation and prognosis prediction tasks for a better performance. In comparison with other deep-learning-based methods, the proposed method achieved the best results on both lesion segmentation (in terms of DSC) and prognosis prediction (in terms of accuracy), as shown in Table 3. Results from our ablation study, as shown in Table 4, further demonstrated the effectiveness of the combination of multi-task learning with joint deep supervision.

It is worth to compare the proposed method with SOTA methods in lymphoma segmentation and outcome prediction. Due to the fact that most of SOTA methods are not open source and that there exist few public data on lymphoma segmentation and outcome prediction available, direct comparison of different methods is difficult. Thus, the comparison results in Table 5 should be interpreted cautiously. Nevertheless, as shown in Table 5, our method achieved the best results on both lymphoma segmentation (in terms of DSC) and outcome prediction (in terms of AUC). One possible explanation why the present approach achieves better results than the existing SOTA methods is attributed to the fact that our approach leverages multi-task learning with deep supervision such that the learned image features of one task

are shared and thereby mutually reinforce the learning of the other task.

Our study has several limitations such as the retrospective design and the modest cohort size. However, with a sample size of 269 patients (89 cases with 2y-EFS and 180 cases with 2y-CR), our cohort size is larger than most of the studies on prognosis prediction at present [6], [7], [26], [27]. Second, we have only trained and validated our model on data from one center, which were acquired using one PET/CT scanner. Whether the proposed model can be generalized to data acquired using other scanners or to data from other centers needs to be further checked. Third, our method combines multi-task learning with deep supervision on both tasks, which leads to the requirement of large GPU memory footprint. In the future, we will explore knowledge distillation [42] to create a lightweight CNN.

VII. CONCLUSION

In this paper we proposed a novel multi-task 3D CNN model for simultaneous lymphoma lesion segmentation and prognosis prediction from baseline FDG-PET images. To reduce overfitting and facilitate network convergence, we further proposed to train our multi-task deep learning model using deep supervision mechanism, which can improve the performance of lesion segmentation and prognosis prediction. After thorough validation, our model may be used to asset the physician as a second opinion while making the final decision. Our future work will focus on creating lightweight CNN with knowledge distillation.

ACKNOWLEDGMENT

(Peng Liu and Miao Zhang contributed equally to this work.)

REFERENCES

[1] E. Roman and A. G. Smith, "Epidemiology of lymphomas," *Histopathology*, vol. 58, no. 1, pp. 4–14, Jan. 2011.

[2] M. J. Maurer, H. Ghesquières, J.-P. Jais, T. E. Witzig, C. Haioun, C. A. Thompson, R. Delarue, I. N. Micallef, F. Peyrade, and W. R. Macon, "Event-free survival at 24 months is a robust end point for disease-related outcome in diffuse large b-cell lymphoma treated with immunochemotherapy," *J. Clin. Oncol.*, vol. 32, no. 10, p. 1066, 2014.

[3] A. Galaznik, C. Reich, G. Klebanov, Y. Khoma, E. Allakhverdiiev, G. Hather, and Y. Shou, "Predicting outcomes in patients with diffuse large b-cell lymphoma treated with standard of care," *Cancer Informat.*, vol. 18, Mar. 2019, Art. no. 1176935119835538.

[4] D. J. Kim, T. Kim, J.-Y. Jeong, J.-C. Jo, W. S. Lee, H.-J. Shin, J. H. Lee, and H. S. Lee, "Poor prognostic impact of high serum ferritin levels in patients with a lower risk of diffuse large b cell lymphoma," *Int. J. Hematol.*, vol. 111, no. 4, pp. 559–566, Apr. 2020.

[5] S. Jemaa, J. Fredrickson, R. A. D. Carano, T. Nielsen, A. de Crespigny, and T. Bengtsson, "Tumor segmentation and feature extraction from whole-body FDG-PET/CT using cascaded 2D and 3D convolutional neural networks," *J. Digit. Imag.*, vol. 33, no. 4, pp. 888–894, Aug. 2020.

[6] Y. Zhang, L. Song, M. Zhao, and K. Hu, "A better prediction of progression-free survival in diffuse large B-cell lymphoma by a prognostic model consisting of baseline TLG and %δSUVmax," *Cancer Med.*, vol. 8, no. 11, pp. 5137–5147, Sep. 2019.

[7] M. Sasanelli, M. Meignan, C. Haioun, A. Berriolo-Riedinger, R.-O. Casasnovas, A. Biggi, A. Gallamini, B. A. Siegel, A. F. Cashen, P. Véra, H. Tilly, A. Versari, and E. Itti, "Pretherapy metabolic tumour volume is an independent predictor of outcome in patients with diffuse large B-cell lymphoma," *Eur. J. Nucl. Med. Mol. Imag.*, vol. 41, no. 11, pp. 2017–2022, Nov. 2014.

[8] H. Senjo, K. Hirata, K. Izumiyama, K. Minauchi, E. Tsukamoto, K. Itoh, M. Kanaya, A. Mori, S. Ota, D. Hashimoto, T. Teshima, and N. Japan Hematology Study Group, "High metabolic heterogeneity on baseline 18FDG-PET/CT scan as a poor prognostic factor for newly diagnosed diffuse large B-cell lymphoma," *Blood Adv.*, vol. 4, no. 10, pp. 2286–2296, May 2020.

[9] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2016, pp. 424–432.

[10] B. Berthon, E. Spezi, P. Galavis, T. Shepherd, A. Apte, M. Hatt, H. Fayad, E. D. Bernardi, C. D. Soffientini, C. R. Schmidtlein, I. El Naqa, R. Jeraj, W. Lu, S. Das, H. Zaidi, O. R. Mawlawi, D. Visvikis, J. A. Lee, and A. S. Kirov, "Toward a standard for the evaluation of PET-auto-segmentation methods following the recommendations of AAPM task group, no. 211: Requirements and implementation," *Med. Phys.*, vol. 44, no. 8, pp. 4098–4111, Aug. 2017.

[11] M. Hatt, J. A. Lee, C. R. Schmidtlein, I. E. Naqa, C. Caldwell, E. D. Bernardi, W. Lu, S. Das, X. Geets, and V. Gregoire, "Classification and evaluation strategies of auto-segmentation approaches for PET: Report of AAPM task group, no. 211," *Med. Phys.*, vol. 44, no. 6, pp. e1–e42, Jun. 2017.

[12] Y. Yu, P. Decazes, J. Lapuyade-Lahorgue, I. Gardin, P. Vera, and S. Ruan, "Semi-automatic lymphoma detection and segmentation using fully conditional random fields," *Computerized Med. Imag. Graph.*, vol. 70, pp. 1–7, Dec. 2018.

[13] H. Hu, P. Decazes, P. Vera, H. Li, and S. Ruan, "Detection and segmentation of lymphomas in 3D PET images via clustering with entropy-based optimization strategy," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, no. 10, pp. 1715–1724, Oct. 2019.

[14] I. J. Hussein, M. A. Burhanuddin, M. A. Mohammed, M. Elhoseny, B. Garcia-Zapirain, M. S. Maashi, and M. S. Maashi, "Fully automatic segmentation of gynaecological abnormality using a new Viola–Jones model," *Comput., Mater. Continua*, vol. 66, no. 3, pp. 3161–3182, 2021.

[15] I. J. Hussein, M. A. Burhanuddin, M. A. Mohammed, N. Benameur, M. S. Maashi, and M. S. Maashi, "Fully-automatic identification of gynaecological abnormality using a new adaptive frequency filter and histogram of oriented gradients (HOG)," *Expert Syst.*, vol. 39, no. 3, Mar. 2022, Art. no. e12789.

[16] S. Li, H. Jiang, H. Li, and Y.-D. Yao, "AW-SDRLSE: Adaptive weighting and scalable distance regularized level set evolution for lymphoma segmentation on PET images," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 4, pp. 1173–1184, Apr. 2021.

[17] E. Grossiord, H. Talbot, N. Passat, M. Meignan, and L. Najman, "Automated 3D lymphoma lesion segmentation from PET/CT characteristics," in *Proc. IEEE 14th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2017, pp. 174–178.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 25. Stateline, NV, USA, Dec. 2012, pp. 1097–1105.

[19] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[20] A. J. Weisman, M. W. Kieler, S. Perlman, M. Hutchings, R. Jeraj, L. Kostakoglu, and T. J. Bradshaw, "Comparison of 11 automated PET segmentation methods in lymphoma," *Phys. Med. Biol.*, vol. 65, no. 23, Dec. 2020, Art. no. 235019.

[21] N. Capobianco, M. Meignan, A.-S. Cottereau, L. Vercellino, L. Sibille, B. Spottiswoode, S. Zuehlsdorff, O. Casasnovas, C. Thieblemont, and I. Buvat, "Deep-learning 18F-FDG uptake classification enables total metabolic tumor volume estimation in diffuse large B-cell lymphoma," *J. Nucl. Med.*, vol. 62, no. 1, pp. 30–36, Jan. 2021.

[22] P. Blanc-Durand, S. Jégou, S. Kanoun, A. Berriolo-Riedinger, C. Bodet-Milin, F. Kraeber-Bodéré, T. Carlier, S. L. Gouill, R.-O. Casasnovas, M. Meignan, and E. Itti, "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. Imag.*, vol. 48, no. 5, pp. 1362–1370, May 2021.

[23] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Dec. 2020.

[24] G. Xu, H. Cao, J. K. Udupa, Y. Tong, and D. A. Torigian, "DiSegNet: A deep dilated convolutional encoder–decoder architecture for lymph node segmentation on PET/CT images," *Computerized Med. Imag. Graph.*, vol. 88, Mar. 2021, Art. no. 101851.

[25] M. Wang, H. Jiang, T. Shi, and Y.-D. Yao, "HD-RDS-UNet: Leveraging spatial-temporal correlation between the decoder feature maps for lymphoma segmentation," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 3, pp. 1116–1127, Mar. 2022.

[26] X. C. Nguyen, W. W. Lee, A. M. Amin, J. S. Eo, S.-M. Bang, J. S. Lee, and S. E. Kim, "Tumor burden assessed by the maximum standardized uptake value and greatest diameter on FDG-PET predicts prognosis in untreated diffuse large B-cell lymphoma," *Nucl. Med. Mol. Imag.*, vol. 44, no. 1, pp. 39–44, Apr. 2010.

[27] N. Aide, C. Fruchart, C. Nganoa, A.-C. Gac, and C. Lasnon, "Baseline 18F-FDG PET radiomic features as predictors of 2-year event-free survival in diffuse large b cell lymphomas treated with immunochemotherapy," *Eur. Radiol.*, vol. 30, no. 8, pp. 4623–4632, Aug. 2020.

[28] X. Zhang, L. Chen, H. Jiang, X. He, L. Feng, M. Ni, M. Ma, J. Wang, T. Zhang, S. Wu, and R. Zhou, "A novel analytic approach for outcome prediction in diffuse large B-cell lymphoma by [18F] FDG PET/CT," *Eur. J. Nucl. Med. Mol. Imag.*, vol. 49, no. 4, pp. 1298–1310, 2022.

[29] J. J. Eertink, T. van de Brug, S. E. Wiegers, G. J. C. Zwezerijnen, E. A. G. Pfaehler, P. J. Lugtenburg, B. van der Holt, H. C. W. de Vet, O. S. Hoekstra, R. Boellaard, and J. M. Zijlstra, "18F-FDG PET baseline radiomics features improve the prediction of treatment outcome in diffuse large B-cell lymphoma," *Eur. J. Nucl. Med. Mol. Imag.*, vol. 49, no. 3, pp. 932–942, Feb. 2022.

[30] L. Sibille, R. Seifert, N. Avramovic, T. Vehren, B. Spottiswoode, S. Zuehlsdorff, and M. Schäfers, "18F-FDG PET/CT uptake classification in lymphoma and lung cancer by using deep convolutional neural networks," *Radiology*, vol. 294, no. 2, pp. 445–452, Feb. 2020.

[31] Z. Liu, S. Wang, J. W. Di Dong, C. Fang, X. Zhou, K. Sun, L. Li, B. Li, M. Wang, and J. Tian, "The applications of radiomics in precision diagnosis and treatment of oncology: Opportunities and challenges," *Theranostics*, vol. 9, no. 5, p. 1303, 2019.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[34] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 472–480.

[35] X. Li, Q. Dou, H. Chen, C.-W. Fu, X. Qi, D. L. Belavỳ, G. Armbrecht, D. Felsenberg, G. Zheng, and P.-A. Heng, "3Dmulti-scale FCN with random modality voxel dropout learning for intervertebral disc localization and segmentation from multi-modality MR images," *Med. Image Anal.*, vol. 45, pp. 41–54, Apr. 2018.

[36] S. Li, Y. Liu, X. Sui, C. Chen, G. Tjio, D. S. W. Ting, and R. S. M. Goh, "Multi-instance multi-scale CNN for medical image classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2019, pp. 531–539.

[37] Y. Zhou, H. Chen, Y. Li, Q. Liu, X. Xu, S. Wang, P.-T. Yap, and D. Shen, "Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 101918.

[38] P. Wang, V. M. Patel, and I. Hacihaliloglu, "Simultaneous segmentation and classification of bone surfaces from ultrasound using a multi-feature guided CNN," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2018, pp. 134–142.

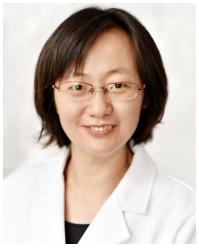
[39] C. Chen, W. Bai, and D. Rueckert, "Multi-task learning for left atrial segmentation on GE-MRI," in *Proc. Int. Workshop Stat. Atlases Comput. Models Heart*. Cham, Switzerland: Springer, 2018, pp. 292–301.

[40] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

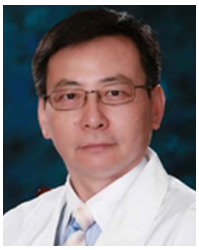
[42] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, pp. 1789–1819, Mar. 2021.

PENG LIU received the master's degree from the University of Chinese Academy of Science (UCAS), in 2018. He is currently pursuing the Ph.D. degree with the Institute of Medical Robotics, Shanghai Jiao Tong University. His research interests include deep learning in medical image analysis and semi-supervised learning.



MIAO ZHANG received the Ph.D. degree from Shanghai Jiao Tong University, in 2009. She is currently working as a Physician with the Department of Nuclear Medicine, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine. Her research interests include PET/CT and PET/MR multimodal molecular imaging, neuroimaging, and deep learning in medical imaging. She is a member of Youth Commission Neurology Group of Chinese Society of Nuclear Medicine (CSNM) and Youth Commission of a member of Shanghai Nuclear Society and Molecular Imaging Committee.

XIAORU GAO (Graduate Student Member, IEEE) received the bachelor's degree from Southern Medical University, in 2020. She is currently pursuing the Ph.D. degree in biomedical engineering with Shanghai Jiao Tong University. Her research interests include medical image registration and deep learning.



BIAO LI received the Ph.D. degree from Shanghai Second Medical University. From 1992 to 1993, he went to Pellegrin Hospital Affiliated to Bordeaux Second University, France, to engage in brain imaging and inflammation imaging research. From 1999 to 2001, he went to the Cancer Biology Laboratory, New England Medical Center, Boston and studied the molecular mechanism of breast cancer. He is currently working as the Director of the Department of Nuclear Medicine, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine as well as a Professor, the Chief Physician, and a Ph.D. Supervision. His research interests include the basic research and clinical application of molecular nuclear medicine. He is currently the Vice Chairperson of the Radiation and Therapy Specialty of Shanghai Nuclear Society, the Deputy Leader of China PET/MR Working Committee, a member of Chinese Society of Nuclear Medicine (CSNM), and the Director of China Nuclear Society.

GUOYAN ZHENG (Member, IEEE) received the Ph.D. degree from the University of Bern, Switzerland, in 2002. He joined the University of Bern, in 2003. Since 2019, he has been a Full Professor at Shanghai Jiao Tong University, China. His research interests include medical image computing, machine learning, computer vision, computer assisted interventions, and medical robotics. Between 2018 and 2022, he was a member of the board of directors of the International Society for Medical Image Computing and Computer Assisted Interventions (MICCAI). He serves as an Associate Editor for IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING and IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS.