Modality-Specific Segmentation Network for Lung Tumor Segmentation in PET-CT Images

Dehui Xiang, *Member, IEEE*, Bin Zhang, Yuxuan Lu, and Shengming Deng

***Abstract*—Lung tumor segmentation in PET-CT images plays an important role to assist physicians in clinical application to accurately diagnose and treat lung cancer. However, it is still a challenging task in medical image processing field. Due to respiration and movement, the lung tumor varies largely in PET images and CT images. Even the two images are almost simultaneously collected and registered, the shape and size of lung tumors in PET-CT images are different from each other. To address these issues, a modality-specific segmentation network (MoSNet) is proposed for lung tumor segmentation in PET-CT images. MoSNet can simultaneously segment the modality-specific lung tumor in PET images and CT images. MoSNet learns a modality-specific representation to describe the inconsistency between PET images and CT images and a modality-fused representation to encode the common feature of lung tumor in PET images and CT images. An adversarial method is proposed to minimize an approximate modality discrepancy through an adversarial objective with respect to a modality discriminator and reserve modality-common representation. This improves the representation power of the network for modality-specific lung tumor segmentation in PET images and CT images. The novelty of MoSNet is its ability to produce a modality-specific map that explicitly quantifies the modality-specific weights for the features in each modality. To demonstrate the superiority of our method, MoSNet is validated in 126 PET-CT images with NSCLC. Experimental results show that MoSNet outperforms state-of-the-art lung tumor segmentation methods.**

***Index Terms*—Lung tumor, PET-CT image, conditional generative adversarial network.**

I. Introduction

LUNG tumor is one of top malignant cancers [1], and it has received great attention from all over the world. Due to increased air pollution and rapid prevalence of tobacco, the number of people with lung cancers increases dramatically, and the mortality rate rises. The incidence and mortality of lung cancer are extremely high, which seriously threatens people's life and health [2]. Positron Emission Tomography (PET) is a non-invasive imaging technology. The radiolabeled glucose analogue 18-fluorodeoxyglucose (FDG) is injected into the human body as a tracer to measure the rate of glucose consumption. Physiological function and biochemical characteristics through the metabolism of specific organs or tissues can be therefore evaluated. PET can be used to detect the metabolism of biological tissues at the molecular and cellular levels. The tumor usually peforms with a higher standardized uptake value (SUV) in PET images; while normal tissue with lower SUV is dark in PET images. Therefore, tumors can be distinguished, and PET images play a very important role in the clinical diagnosis and treatment [3]. However, the disadvantage of PET images is that the spatial resolution is relatively low, and the boundary of lesions often appears blurred. CT images can provide information on the anatomical structure of various organs or tissues, but they cannot show the presence of the metabolic information. Therefore, it is difficult to distinguish abnormal and normal organs or tissues. With the introduction of multi-modality imaging technologies, PET-CT scanners can provide paired FDG-PET and CT images, which make it possible to simultaneously acquire both functional and anatomical images [4]. Therefore, PET-CT imaging is often used in clinical application to diagnose lung cancer, such as non-small cell lung cancer (NSCLC), which is the most common type of lung cancer. PET-CT images have become common objects for lung tumor segmentation and cancer assessment, and gained a lot of attention in the field of image processing.

Although PET-CT images have been widely used in clinic, lung tumor segmentation is still a challenging task in medical image processing field [5]–[9]. As can be seen in Fig. 1, the main difficulties of lung tumor segmentation in PET-CT images include the following four points. First, discrepancy often occurs when the lung tumor is visualized in PET images and CT images. Intensities of the lung tumor vary largely in PET images and CT images. CT provides the anatomical localization of the tumor while PET provides the function and the glucose metabolism. The curves in Fig. 1 are the contours of the lung tumors, the blue curves are ground truth of the CT image, and the green curves are ground truth of the PET image. The two corresponding curves are not consistent in the two modality images from the same patients. As studied in previous works [10]–[12], tumor position between PET and CT is different even using the same protocol at different respiration levels. The shape and size of lung tumors in PET-CT images are different from each other since delay maybe occur in the process of capturing the two modality images

Manuscript received 30 December 2021; revised 29 March 2022 and 12 May 2022; accepted 22 June 2022. Date of publication 27 June 2022; date of current version 7 March 2023. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61971298, and in part by the National Key R&D Program of China under Grant 2018YFA0701700. *(Dehui Xiang and Bin Zhang contributed equally to this work.) (Corresponding author: Dehui Xiang.)*

Dehui Xiang and Yuxuan Lu are with the School of Electronic and Information Engineering, Soochow University, Jiangsu 215006, China (e-mail: xiangdehui@suda.edu.cn; 20194228016@stu.suda.edu.cn).

Bin Zhang and Shengming Deng are with the First Affiliated Hospital of Soochow University, Suzhou 215006, China (e-mail: zbnuclmd@126.com; dshming@163.com).

Digital Object Identifier 10.1109/JBHI.2022.3186275

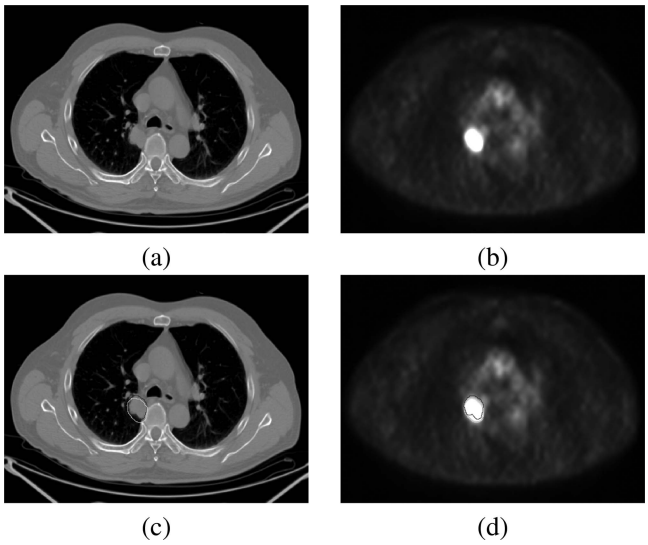

Fig. 1. Challenges in lung tumor segmentation. The blue curves are ground truth of the CT image, and the green curves are ground truth of the PET image. (a) Original CT image; (b) Original PET image; (c) CT image with ground truth of the CT image (a) and PET image (b); (d) PET image with ground truth of the CT image (a) and PET image (b).

and the patients may not comply with standard breath-hold protocol. Image registration is not perfect after the PET-CT scanner produces two modality images. Second, the boundary of the lung tumor is visually blurred in CT images, and the contrast between surrounding tissues and the tumor is low, which make it very difficult to distinguish its boundary. Third, there are several neighboring organs, e.g. liver, heart, and muscles, and they share the similar intensities both in CT images and PET images between the tumor and its neighboring organs [12], [13]. Fourth, the shape and location of the tumor have great anatomical differences between different patients, especially the tumor may exist anywhere in the thorax. Therefore, to address the above issues, an effective segmentation framework is designed for robust modality-specific lung tumor segmentation.

The combination of PET and CT has improved the diagnostic capability of the lung tumor in clinic practice [4], [14], [15]. Many previous papers also have indicated that the integration of PET and CT information can produce more accurate tumor volume [6], [7], [9], [13], [16]–[19]. Our previous work [13], [20] in PET-CT images and CT images show promising for lung and tumor segmentation, but these conventional methods need users' interaction or initialization. Convolutional neural networks (CNNs) can reduce a lot of efforts in preprocessing steps and make it more automatic to detect, classify and segment images.

Considering the complementarity and inconsistency between PET images and CT images, our aim is to fuse the complementary information in the two modality images for automatic lung tumor segmentation, and meanwhile, to preserve modality-specific features of PET images and CT images. In particular, we focus on our method can simultaneously obtain the corresponding lung tumor segmentations of PET images and CT images. We follow Kumar's work [6] to fuse complementary anatomical and functional information of the lung tumor from PET-CT images in the image intensity varying manner. As reported by Ligtenberg *et al.* [10], PET-based clinical target volumes (CTV) were significantly smaller compared to CT-based CTVs. It shows that it is necessary to provide a modality-specific target definition of the tumor. Therefore, we focus on modality-specific features that improve modality-specific segmentation since CT images depict the lung tumor across multiple anatomy and PET images depict the lung tumor across multiple locations of function and the glucose metabolism.

Therefore, a modality-specific segmentation network (MoSNet) is proposed for modality-specific lung tumor segmentation in PET-CT images. MoSNet learns a modality-specific representation to describe the inconsistency between PET images and CT images and a modality-fused representation to encode the common feature of lung tumor in PET images and CT images. An adversarial method is proposed to minimize an approximate modality discrepancy through an adversarial objective with respect to a modality discriminator and reserve modality-common representation. This improves the representation power of the network for modality-specific lung tumor segmentation in PET images and CT images. The novelty of MoSNet is its ability to produce a modality-specific map that explicitly quantifies the modality-specific weights for the features in each modality. This is in contrast to CNNs that use produce a single map for CT images or PET images [6]–[9], [18], [19], [21]. MoSNet is intended as a modality-specific approach for integrating PET and CT information and keeping modality-fused features to obtain modality-specific segmentation. To demonstrate the efficacy of our method, we conduct experimental comparisons with state-of-the-art lung tumor segmentation methods on PET-CT images with NSCLC.

II. Related Work

Accurate segmentation of lung tumors plays a very important role in clinical diagnosis and treatment. A large number of lung tumor studies have been reported to improve the accuracy of lung tumor segmentation. Overall, current lung tumor segmentation methods can be divided into traditional methods and deep learning methods.

Traditionally, SUV was often used for lung tumor segmentation in PET images. A tumor would be considered as a malignant tumor in clinical diagnosis when SUV is higher than a constant. Therefore, a large number of lung tumor segmentation algorithms are based on threshold values. Erdi $et\ al.$ [22] used a fixed threshold to predict the true lesion volume for lesions larger than 4 mL based only on the source–to–background value from PET images. However, the fixed threshold method produced overestimation of the volume by an amount that depended on the source–to–background ratio for smaller volumes. Jentzen $et\ al.$ [23] proposed an iterative thresholding method, and their method performed well only in the visible area of PET images. Nehmeh $et\ al.$ [24] developed an iterative method to estimate threshold value of tumor segmentation based on Monte Carlo simulation. They described the correlation between lesion volume and the

corresponding optimal threshold so that the optimal threshold value could be determined. However, threshold methods are often sensitive to uneven grayscale distribution of tumors in PET images. The gray values of tumors are similar to those of the liver, heart, and spine. Under-segmentation or over-segmentation usually occurs, resulting in low true positive or high false positive in the segmentation results. Tumor features had further studied to extract from PET images to segment tumors. Geets *et al.* [25] proposed a gradient-based tumor segmentation method on PET images, and they combined gradient intensity estimation with watershed transformation and hierarchical clustering analysis. Belhassen *et al.* [26] proposed a fuzzy C-Means clustering algorithm to cope with noisy and low resolution PET images. Nonlinear anisotropic diffusion filtering was first used to smooth PET images, and then fuzzy C-Means clustering and spatial information were combined. In recent years, more segmentation algorithms have been proposed to segment tumors. Cherry *et al.* [27] proposed a graph cut method to improve the segmentation of PET lung tumors. This energy function of the graph cut method was based on the tumor voxel in the PET image, and combined with a SUV function. The monotonic declining characteristic was supposed to solve the problem of tumor heterogeneity and to boost the segmentation of tumors from adjacent structures also with high FDG uptake. PET and CT images can complement each other. Jafar *et al.* [28] proposed a tumor segmentation method for CT and PET images. The method was designed to find the optimal threshold value to extract lung from the three-dimensional volume, and a multi-threshold algorithm was used to detect all suspicious PET and CT images. Fuzzy clustering algorithm was also used to reduce false positive. Guo *et al.* [29] proposed a reliable method for automatic lung tumor segmentation on PET and CT images based on the fuzzy Markov random field model. The method was a combination of PET and CT image information through the observation of posterior probability distribution. Soltani-Nabipour *et al.* [30] used improved region growing algorithm in CT images. Vijh *et al.* [31] used marker-controlled watershed and support vctor machine to segment and classify lung tumor. Han *et al.* [32] also proposed a Markov random field based segmentation of the image pair with a regularized term that penalizes the segmentation difference between PET and CT to concurrently segment tumor from both modalities. Song *et al.* [16] constructed two sub-graphs for the segmentation of the PET and the CT images. An adaptive context cost was proposed by adding context arcs to achieve consistent results in two modalities. Ulaş *et al.* [17] developed a graph-based interactive segmentation method, and proposed random walk to segment PET, PET-CT, MRI-PET, and MRI-PET-CT images to obtain the global optimal contour of tumors [33]. We [13] used a random walk algorithm to obtain the initial contour of lung tumors, and proposed a graph cut algorithm with a joint segmentation energy function for PET and CT images, and lung tumors were obtained by minimizing the energy function. However, these methods often needed users' interaction.

Deep learning methods have already been demonstrated advantages in disease diagnosis on medical images, such as abnormality detection and segmentation due to their huge power in extracting useful information from large amount of data. Convolutional neural networks (CNNs) are one of popular deep learning methods to address the common semantic segmentation or detection tasks. CNN architecture such as U-Net appears to be a encode-decode structure to hierarchically fuse low-level and high-level features with the combination of convolution and deconvolution layers. Jiang *et al.* [34] proposed multiple resolution residually connected network to simultaneouslycombine features across multiple image resolution and feature levels through residual connections to detect and segment the lung tumors in CT images. Zhao *et al.* [21] proposed a multi-modality segmentation method based on a 3D fully convolutional neural network (FCN) to use both PET and CT information simultaneously for tumor segmentation. Xu *et al.* [18] assembled PET and CT into two channels of combined images and cascaded two V-Nets to form a W-Net architecture to improve the segmentation to bone-specifc lesions. Zhong *et al.* [19] used two 3D-UNets to respectively train on the preprocessed CT image and PET image to obtain the coarse segmentation results of lung tumors, and then further adopted a joint graph segmentation method based on potential label consistency between of PET and CT bimodal images to refine initial segmentation. Kumar *et al.* [6] proposed a co-learning feature fusion CNN model to fuse complementary information for PET-CT images. The model encoded modality-specific features to generate a spatially varying fusion map that quantifies the relative importance of each modality's features across different spatial locations and to obtain a representation of the complementary multi-modality information at different locations. Lu *et al.* [8] constructed a neural network architecture for auto-segmenting tumors by leveraging a 14-layer U-Net model with two blocks of a VGG19 encoder pre-trained with ImageNet. They then imported a DropBlock technique to replace the normal regularization dropout method to help U-Net efficiently avoid overfitting. Their method achieved a relatively competitive performance in PET images on tumor segmentation. Li *et al.* [7] also designed a 3D FCN to produce a probability map from the CT image and roughly segmented the tumor from its surrounding soft tissues. Hu *et al.* [35] proposed a hybrid attention mechanism and densely connected convolutional networks to segment lumg tumor in CT images. Pang *et al.* [36] introduced CTumorGAN to segment lung tumor in CT images. Gan *et al.* [37] combined 2D dense connection CNN and 3D V-Net to segment lung tumor in CT images. Tyagi *et al.* [38] used 3D U-Net with deformable convolution blocks to segment lung tumor in CT images. Fu *et al.* [9] introduced a multimodal spatial attention module (MRRN) that automatically learned to emphasize spatial regions related to tumors and suppress normal regions with physiologic high-uptake in PET images. The spatial attention maps were subsequently employed in a CNN for segmentation of areas with higher tumor likelihood in the CT images. Li *et al.* [39] proposed the cycle-consistent image conditional variational autoencoder and the Res-Unet to segment lung tumor on multi-modal MRI images. Dutande *et al.* [40] introduced two deep residual separable convolutional neural networks to lung tumor from CT images. Zhao *et al.* [41] used a distraction-Ssensitive U-Net to segment lung tumor in CT images. In recent years, domain adaptation has attracted

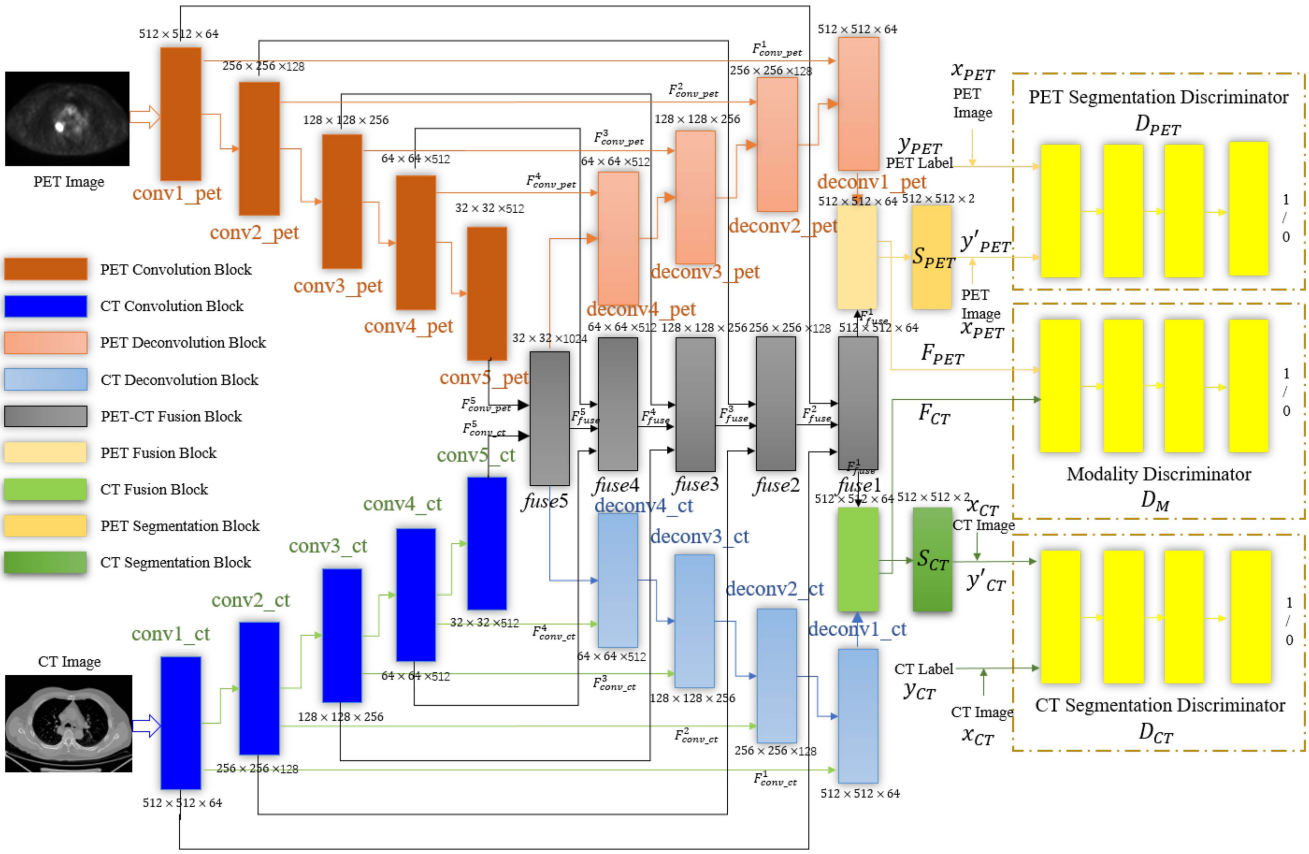

Fig. 2. The architecture of MoSNet.

researchers' attention to transfer knowledge from a well-labeled and related source domain to assist the target domain learning [42], [43]. Domain transfer networks have been proposed to extract domain shift features by minimizing some measure of domain gap such as maximum mean discrepancy (MMD) [43]–[45] or correlation distances [46], [47]. GANs [48] have also been used to domain transfer tasks [49]. MMD is the norm of the difference between two domain features. Feature discrepancy across domains was evaluated by MMD embedded in a reproducing kernel Hilbert space [43]–[45]. Correlation distance computed the mean and covariance of the two distributions [46], [47]. Shallow fully connected network was also designed as a domain classifier to reduce domain discrepancy [50], [51]. Domain shift was treated as a binary classification problem. A gradient reversal layer (GRL) was proposed to change the sign of the gradient from the subsequent level for backpropagation. GANs have also been explicitly used to transfer a sample in one domain to an analog sample in another domain. The domain transfer network was proposed to learn a generative function that maps an input sample from source domain to the target domain with a compound loss function. The discriminative classifier distinguishes between fake images drawn from the generator and true images from the training data. These domain adaptation methods are promising techniques for cross-domain tasks and often used to unsupervised visual domain adaptation. Inspired by these work, these adversarial methods are used to minimize an approximate modality discrepancy between PET-CT images.

III. Method

Fig. 2 shows the whole architecture of lung tumor segmentation network. Multiple modality-specific encoder-decoder branches are integrated into the conditional Generative Adversarial Network (cGAN) framework, and a novel PET-CT convolution neural network is therefore proposed for lung tumor segmentation. The dual-stream encoder EN_{PET}, EN_{CT}, is respectively designed to generate different level modality-specific representations for PET images and CT images. One decoder branch DE_{fuse} is designed to fuse the multiple modality representations of PET images and CT images. A modality-specific decoder branch DE_{PET} is designed to integrate different level modality-specific representations of PET images. Another modality-specific decoder branch DE_{CT} is also designed to integrate different level modality-specific representations of CT images. The two modality-specific decoder branches DE_{PET} and DE_{CT} are respectively further fused with DE_{fuse} to generate fused modality-specific features F_{PET} and F_{CT}. PET segmentation discriminator D_{PET} and CT segmentation discriminator D_{CT}

are used to distinguish the lung tumor segmentation. A modality discriminator D_M is proposed to distinguish modality-specific features F_{PET} and F_{CT}. The structure of the proposed model is described below.

A. Modality-Specific Encoder

The PET encoder EN_{PET} includes five PET convolution blocks: $conv1_{pet}$, $conv2_{pet}$, $conv3_{pet}$, $conv4_{pet}$, $conv5_{pet}$, where the size of input features is downsampled as 1/2 in width and height of the former convolution block features except $conv1_{pet}$. The down-sampling layers of the convolution block adopts a maxpooling layer with the stride of 2 and the kernel size of 2×2. Each convolution block includes convolution (the size of convolution kernels is 3×3, and the stride is 1) and batch normalization. The input of the activation function ReLU is connected to the output of each batch normalization layer. The convolution block is defined as

$$DCovv_{3\times3}\left(F_{in}\right) = Cov_{3\times3}\left(Cov_{3\times3}\left(MP\left(F_{in}\right)\right)\right), \quad (1)$$

$$Cov_{3\times3}\left(F_{in}\right) = ReLU\left(BN\left(conv_{3\times3}\left(F_{in}\right)\right)\right), \quad (2)$$

where F_{in} is the input feature, MP is a maxpooling operation and $DCov$ is a double convolutional block. Kindly note that input of $conv1_{pet}$ does not need to be computed with a maxpooling layer. $ReLU$ is used as the activation function to eliminate negative feature values, linearly emphasize regions of high saliency with large values, such that the generator can restrain non-tumor features and emphasize tumor features. BN is batch normalization and $conv_{3\times3}$ is the 3×3 convolution operation. The CT encoder EN_{CT} also includes five CT convolution blocks: $conv1_{ct}$, $conv2_{ct}$, $conv3_{ct}$, $conv4_{ct}$, $conv5_{ct}$ and the five convolution blocks are the same as those of PET.

B. Modality-Specific Fusion

The fusion decoder branch DE_{fuse} includes five PET-CT fusion blocks: $fuse1$, $fuse2$, $fuse3$, $fuse4$, $fuse5$. The PET-CT fusion block connects PET feature and CT feature. Each level PET-CT feature fusion includes a concatenation operation and a double convolution block without the maxpooling operation in (1). The PET-CT feature fusion in lth level is defined as

$$F^l_{fuse} = DCov\left(concat\left(F^l_{conv\_pet}, F^l_{conv\_ct}\right)\right), \quad (3)$$

where $concat$ is a concatenation operation. $F^l_{conv\_pet}$ denotes the output feature of $convl_{pet}$ in lth level and $F^l_{conv\_ct}$ denotes the output feature of $convl_{ct}$ in lth level. Level lth feature is then up-sampled using a 2×2 transpose convolution with a stride of 2×2. Level lth feature with the transpose convolution is concatenated with F^{l-1}_{fuse} and followed by a double convolution block without the maxpooling operation in (1).

The PET modality-specific decoder branch DE_{PET} includes four PET deconvolution blocks: $deconv1_{pet}$, $deconv2_{pet}$, $deconv3_{pet}$, $deconv4_{pet}$. Level lth feature of PET is then up-sampled using a 2×2 transpose convolution with a stride of 2×2. Level lth feature with the transpose convolution is concatenated with PET feature F^{l-1}_{pet} and followed by a double convolution block without the maxpooling operation in (1). The CT modality-specific decoder branch DE_{CT} also includes four CT deconvolution blocks: $deconv1_{ct}$, $deconv2_{ct}$, $deconv3_{ct}$, $deconv4_{ct}$ and the four deconvolution blocks are the same as those of PET.

Final modality-specific fusion includes a PET fusion block and a CT fusion block. The PET fusion block further fuses the output features of $fuse1$ and $deconv1_{pet}$. The output features of $fuse1$ and $deconv1_{pet}$ are concatenated and fused with a double convolution block without the maxpooling operation in (1). The feature is fed to PET segmentation block (a 3×3 convolution operation) for generating the segmentation map of the PET image. The segmentation map of the CT image is also generated as that of the PET image.

C. Discriminator

For cGAN segmentation network, the PET segmentation discriminator, the CT segmentation discriminator and the modality discriminator D_M contain three convolutional blocks including a 4×4 convolutional layer, batch normalization and LeakyReLU, and a 1×1 convolutional layer. LeakyReLU is used as the activation function to linearly emphasize regions of high saliency with large values but linearly scale down the negative feature values, such that the discriminators can distinguish true/fake or PET/CT features more accurately. The segmented image or the label image are concatenated with the original image and then fed to the segmentation discriminator. The segmentation discriminator judges to be false when the segmented image is fed and judges to be true when the manually labeled image is fed. The modality discriminator judges to be false when the modality-specific feature F_{CT} is fed and judges to be true when the modality-specific feature F_{PET} is fed.

D. Modality-Specific Segmentation

In our multi-modality image segmentation, PET images and CT images are respectively labeled. x_{PET}, y_{PET} denote a PET image and its corresponding label. Similarly, x_{CT}, y_{CT} denote a CT image and its corresponding label. PET and CT image pairs have the same classes from the same patient, but there exists a domain shift between their feature distributions. In this work, we propose a novel framework for modality-specific segmentation, allowing us to effectively integrate PET and CT information and keep modality-fused features to segment the modality-specific lung tumor.

Most existing researches in PET-CT image segmentation are devoted to segment tumors in one modal image without or with guidance of the other modal image [6]–[9], [18], [19], [34], only considering one modal image segmentation. However, these tumor segmentation methods did not include modality discrepancy, which is more practical and challenging in clinical applications. Thus, a reasonable consideration is that, both modality-common and modality-specific features should be simultaneously learned in convolutional fusion stage to effectively model complex data from different modalities. Meanwhile, the modality inconsistency is explicitly minimized with feature alignment after the modality-specific layers: PET fusion block and CT fusion block.

The proposed lung tumor segmentation network is embedded into the cGAN framework [52], as shown in Fig. 2. The cGAN framework learns a mapping S from PET-CT images to two segmentation probabilities for the PET image and the CT image. The PET segmentation discriminator D_{PET} classifies concatenated pairs of PET image and prediction as being real or fake constrained by the adversarial loss, which is calculated from the discriminator D_{PET} to penalize the generator S. D_{CT} classifies concatenated pairs of CT image and prediction as being real or fake constrained by the adversarial loss, which is calculated from the discriminator D_{CT} to penalize the generator S. Given a PET-CT image pair x_{PET}, x_{CT}, and the corresponding manually annotated label pair y_{PET}, y_{CT}, the adversarial loss in cGAN is used to match the distribution of images to that of the target distribution, and it can be expressed as,

$$\begin{aligned} &L_{GAN}^{S,D_{PET},D_{CT}}\left(x_{PET},x_{CT}\right) \\ &= E\left[\log D_{PET}\left(y_{PET},x_{PET}\right)\right] \\ &\quad + E\left[\log\left(1-D_{PET}\left(S_{PET}\left(x_{PET},x_{CT}\right),x_{PET}\right)\right)\right] \\ &\quad + E\left[\log D_{CT}\left(y_{CT},x_{CT}\right)\right] \\ &\quad + E\left[\log\left(1-D_{CT}\left(S_{CT}\left(x_{PET},x_{CT}\right),x_{CT}\right)\right)\right]. \end{aligned} \tag{4}$$

where E is the average function. A pixel-wise loss term L_1 is used to penalize pixel-wise segmentation errors and bring the PET prediction $y_{PET}{}' = S_{PET}(x_{PET}, x_{CT})$ from the generator S ($y_{CT}{}' = S_{CT}(x_{PET}, x_{CT})$ denotes the CT prediction) closer to ground truth y_{PET} and stabilize GAN training,

$$L_1^S\left(x_{PET},y_{PET}\right) = E\left[\left\|y_{PET}-y_{PET}{}'\right\|_1\right], \tag{5}$$

The binary cross-entropy function is used in the segmentation network to classify the foreground and background of each pixel as

$$\begin{aligned} L_{bce}^S\left(y_{PET},y_{PET}{}'\right) = &-\frac{1}{H\times W}\sum_{i=1}^{H\times W}\left(y_{PET_i}\log y_{PET_i}{}'\right., \\ &\left.+\left(1-y_{PET_i}\right)\log\left(1-y_{PET_i}{}'\right)\right), \end{aligned} \tag{6}$$

where $y_{PET_i}{}' \in [0,1]$ is the ith pixel in the PET prediction of $y_{PET}{}'$, $y_{PET_i} \in [0,1]$ is the ith pixel in ground truth y_{PET}. H and W are the height and width of the image, respectively. The binary cross-entropy loss L_{bce}^S is commonly used in classification tasks and may ignore the segmentation integrity of the image level. Therefore, the dice loss be also introduced to optimize our proposed network.

$$L_{dice}^S\left(y_{PET},y_{PET}{}'\right) = 1-\frac{2\sum_{i=1}^{H\times W}\left(y_{PET_i}y_{PET_i}{}'\right)}{\sum_{i=1}^{H\times W}y_{PET_i}+\sum_{i=1}^{H\times W}y_{PET_i}{}'}. \tag{7}$$

$L_1^S(x_{CT}, y_{CT})$, $L_{bce}^S(y_{CT}, y_{CT}{}')$ and $L_{dice}^S(y_{PET}, y_{PET}{}')$ are also defined for CT prediction as those of PET prediction.

Since there exists a certain correlation between PET and CT images, the two modality images should share part of the network parameters. Meanwhile, the PET and CT images are distributed differently, so the modality-specific network with supervised information should be proposed to extract features, which are only sensitive to the corresponding modal images. Therefore, the segmentation network S learns a modality-fused representation to encode the common features of lung tumor in PET images and CT images, and a modality-specific representation to desribe the inconsistency beween PET images and CT images. A modality discriminator D_M is used to minimize PET and CT representation distances classifies whether a representation is drawn from the PET image or the CT image. D_M is optimized according to a supervised loss as

$$\begin{aligned} L_{GAN}^{S,D_M}\left(x_{PET},x_{CT}\right) = &E\left[\log D_M\left(F_{PET}\right)\right], \\ &+E\left[\log\left(1-D_M\left(F_{CT}\right)\right)\right], \end{aligned} \tag{8}$$

where F_{PET} denotes PET-specific feature computed by the PET fusion block, and F_{CT} denotes CT-specific feature computed by the PET fusion block in the segmentation network S. There are other different possible choices of adversarial loss functions, such as MMD [43]–[45], the gradient reversal layer [50], [51], maximum classifier discrepancy (MCD) [53], margin disparity discrepancy (MDD) [54].

Based on (4) to (8), the full objective function for cGAN-based segmentation can be expressed as,

$$\begin{aligned} &L^{S,D_{PET},D_{CT},D_M}\left(x_{PET},y_{PET}\right) \\ &= L_{GAN}^{S,D_{PET},D_{CT}}\left(x_{PET},x_{CT}\right) \\ &\quad + \lambda_1 L_1^S\left(x_{PET},y_{PET}\right) + \lambda_1 L_1^S\left(x_{CT},y_{CT}\right) \\ &\quad + \lambda_{bce} L_{bce}^S\left(y_{PET},y_{PET}{}'\right) + \lambda_{bce} L_{bce}^S\left(y_{CT},y_{PET}{}'\right) \\ &\quad + \lambda_{dice} L_{dice}^S\left(y_{PET},y_{PET}{}'\right) + \lambda_{dice} L_{dice}^S\left(y_{CT},y_{CT}{}'\right) \\ &\quad + L_{GAN}^{S,D_M}\left(x_{PET},x_{CT}\right), \end{aligned} \tag{9}$$

where λ_1 is the weight of the L_1 loss, λ_{bce} is the weight of the L_{bce}^S loss, λ_{dice} is the weight of the L_{dice}^S loss.

To this end, we integrate all the components and obtain the following overall objective of MoSNet, in which the segmentation network S and the three discriminators: PET segmentation discriminator D_{PET}, CT segmentation discriminator D_{CT} and the modality discriminator D_M, play a min-max game in respectively minimizing and maximizing the objective function as

$$\arg\max_{D_{PET},D_{CT},D_M}\min_S L^{S,D_{PET},D_{CT},D_M}. \tag{10}$$

IV. Experiments

In this section, we test the performance of the modality-specific lung tumor segmentation framework by conducting extensive evaluations on our benchmark dataset. Additional details about experiments and results are reported as following.

A. Data

Our dataset comprises 126 FDG PET-CT scans of patients with biopsy-proven NSCLC. The images are acquired by a GE Discovery ST16 PET-CT scanner in the first affiliated hospital of Soochow University and adhered to the tenets of the Declaration of Helsinki. Each image comprises one CT volume and

one PET volume. The two volumes are reconstructed with the same number of slices. The CT resolution is 512×512 pixels at 0.98 mm×0.98 mm, the PET resolution is 128 ×128 pixels at 5.47 mm × 5.47 mm, with a slice thickness and an interslice distance of 3.27 mm. Images contain between 1 to 3 tumors in the thorax. PET images are rescaled to 512 × 512 in axial view so that the PET-CT image pair shares the same coordinate space. Tumors are manually annotated by an experienced oncologic nuclear imaging expert using ITKsnap [55]. Lung tumors in PET images and CT images are respectively annotated with the consideration of the other modality. Only small number of 2D thorax slices contain the lung tumors, many previous researches only trained and tested their models with the 2D thorax slices containing tumors [6], [9], [34]. In our experiments, we do not exclude slices without tumor pixels in ground truth. 126 FDG PET-CT scans are divided into two groups: 62 and 64 as training and test sets for a 2-fold cross validation evaluation. The first group contains 3460 2D thorax PET-CT slice pairs and the second contains 3525 2D thorax PET-CT slice pairs.

B. Implementation Details

All models are built in pytorch. The workstation is with a NVIDIA GeForce RTX 3090 GPU with 24 G memory. The batch size is set to 1 and 110 epochs are trained. Since the proposed model is implemented in cGAN framework and the sizes of the lung tumors in CT images and PET images are often small, Adam optimizer is used to stably optimize the proposed model and the initial learning rate is 0.0002. The learning rate is not changed in the first 10 epochs and then linearly attenuated in the rest 100 epochs. The size of all PET-CT images is 512 × 512. λ_1 is set to 100. λ_{bce} and λ_{dice} are set to 1. The total training time of MoSNet is about 230 hours, and the test time of each PET-CT volume is about 30 seconds.

C. Evaluation Metrics

In order to quantitatively evaluate the algorithms, four commonly used evaluation indicators in medical image segmentation are used as the evaluation criteria for experimental results of PET volumes and CT volumes: Dice similarity coefficient (DSC), intersection over union (IoU), precision and recall. DSC and IoU are usually used to measure the similarity between the network segmentation results and ground truth as

$$Dice = \frac{2TP}{FP + 2TP + FN}, \tag{11}$$

$$IoU = \frac{TP}{FP + TP + FN}, \tag{12}$$

where TP represents the number of true positives, FP represents the number of false positives and FN represents the number of false negatives. Precision and recall are defined as

$$Precision = \frac{TP}{FP + TP}, \tag{13}$$

$$Recall = \frac{TP}{TP + FN}. \tag{14}$$

Paired t-test to Dice is conducted to compare the difference in segmentation results between our method and related methods, and $p < 0.05$ is considered statistically significant.

D. Ablation Experiments

Different combinations and adversarial loss functions are tested in our modality-specific lung tumor segmentation framework to determine the contributions of each block and adversarial loss to the segmentation performance. A comparison of the improvement on tumor segmentation performance is presented in Table I. The configurations of ablation experiments are as follows,

- M1 Baseline+L_1^S+w/oNL: The baseline segmentation architecture is an encoder-decoder network, which includes the modality-specific encoder, PET-CT fusion blocks, PET segmentation block and CT segmentation block. PET segmentation block and CT segmentation block directly connect to $fuse1$ without the PET fusion block and the CT fusion block. PET segmentation discriminator D_{PET} and CT segmentation discriminator D_{CT} are also used to construct the cGAN framework. The loss is L_1^S but normal thorax PET-CT slice pairs (NL) without the lung tumor are not used. The first group contains 895 PET-CT slice pairs with lung tumors and the second group contains 840 PET-CT slice pairs. All the PET-CT slice pairs including normal thorax PET-CT slice pairs are segmented in the test stage.
- M2 Baseline+L_1^S+L_{bce}^S+w/oNL: The baseline segmentation architecture, L_1^S and L_{bce}^S are used, but normal thorax PET-CT slice pairs (NL) are not used.
- M3 Baseline+L_1^S+L_{bce}^S+L_{dice}^S+w/oNL: The baseline segmentation architecture, L_1^S, L_{bce}^S and L_{dice}^S are used, but normal thorax PET-CT slice pairs (NL) are not used.
- M4 Baseline+L_1^S+L_{bce}^S+L_{dice}^S: The baseline segmentation architecture, L_1^S, L_{bce}^S and L_{dice}^S are used, and normal thorax PET-CT slice pairs (NL) are used.
- M5 M4+Mo: The extension of M4. Two decoders and two fusion blocks are added to the baseline segmentation architecture. The CT modality-specific decoder branch and the PET modality-specific (Mo) decoder branch are added. The PET fusion block and the CT fusion block are also added.
- M6 M5+Mo+$L^{S,D_{MMD}}$: The extension of M5. Modality-specific feature alignment is based on the MMD criterion $L^{S,D_{MMD}}$ [43]–[45] with M5 in the reproducing kernel Hilbert space.
- M7 M5+Mo+$L^{S,D_{MMD_P}}$: The extension of M5. Modality-specific segmentation alignment is based on the MMD criterion $L^{S,D_{MMD_P}}$ [43]–[45] with M5 in the reproducing kernel Hilbert space.
- M8 M5+Mo+$L^{S,D_{GRL}}$: The extension of M5. Modality-specific feature alignment is based on the GRL $L^{S,D_{GRL}}$ [50], [51]. Due to the limitation of the

TABLE I
LUNG TUMOR SEGMENTATION RESULTS OF CT IMAGES AND PET IMAGES IN ABLATION EXPERIMENTS (MEAN ± STANDARD DEVIATION)

CT	Precision(%)	Recall(%)	IoU(%)	Dice(%)	p
PET	63.56±19.34	81.53±17.54	54.19±16.37	68.69±15.28	-
Baseline+L_1^S+w/oNL (M1)	47.11±29	73.81±18.99	39.17±24.31	51.74±26.41	2.76×10^{-20}
Baseline+L_1^S+L_{bce}^S+w/oNL (M2)	49.65±28.56	72.5±22.46	41.21±24.63	53.77±26.72	2.32×10^{-19}
Baseline+L_1^S+L_{bce}^S+L_{dice}^S+w/oNL (M3)	58.8±28.75	72.15±21.81	46.11±24.55	58.89±25.52	1.23×10^{-14}
Baseline+L_1^S+L_{bce}^S+L_{dice}^S (M4)	81.28±23.1	70.54±25.04	60.37±24.69	71.68±23.79	0.016
M4+Mo (M5)	74.3±25.15	75.89±18.43	59.49±22.81	71.53±21.75	0.01
M5+Mo+$L^{S,D_{MMD}}$ (M6)	53.23±30.02	76.55±19.42	44.77±26.01	56.97±27.53	7.25×10^{-14}
M5+Mo+$L^{S,D_{MMD_p}}$ (M7)	78.02±26.05	64.17±29.26	54.99±26.65	66.31±27.43	2.13×10^{-6}
M5+Mo+$L^{S,D_{GRL}}$ (M8)	76.65±23	72.5±23.93	58.97±22.99	71.01±22.37	8.68×10^{-4}
M5+Mo+L_{GAN}^{S,D_p} (M9)	83.9±21.01	67.57±25.02	59.51±23.04	71.42±22.48	0.012
M5+Mo+L_{GAN}^{S,D_M} (M10, Our method)	82.34±19.33	74.63±21.84	63.76±20.99	75.52±18.65	
PET	Precision(%)	Recall(%)	IoU(%)	Dice(%)	p
Baseline+L_1^S+w/oNL (M1)	54.82±29.52	82.5±17.61	47.42±25.36	60.02±25.36	3.88×10^{-15}
Baseline+L_1^S+L_{bce}^S+w/oNL (M2)	61.15±30.36	80.75±18.59	52.03±25.55	64.21±25.53	6.92×10^{-11}
Baseline+L_1^S+L_{bce}^S+L_{dice}^S+w/oNL (M3)	61.03±28.03	80.46±18.64	50.92±23.77	63.82±23.52	3.56×10^{-12}
Baseline+L_1^S+L_{bce}^S+L_{dice}^S (M4)	83.76±21.28	74.51±23.48	62.36±22.66	73.96±20.72	0.01
M4+Mo (M5)	78.25±23.93	78.49±21.81	62.12±22.8	73.73±20.95	8.51×10^{-3}
M5+Mo+$L^{S,D_{MMD}}$ (M6)	54.77±29.82	82.12±20.01	47.15±25.83	59.45±26.73	1.05×10^{-15}
M5+Mo+$L^{S,D_{MMD_p}}$ (M7)	80.85±24.74	66.61±30.19	55.93±27.46	66.9±27.83	1.29×10^{-6}
M5+Mo+$L^{S,D_{GRL}}$ (M8)	81.5±22.75	77.6±24.38	64.58±23.1	75.44±22	0.13
M5+Mo+L_{GAN}^{S,D_p} (M9)	85.96±20.7	75.71±23.67	65.18±21.93	76.27±20.31	0.28
M5+Mo+L_{GAN}^{S,D_M} (M10, Our method)	82.55±20.9	79.63±18.85	66.26±19.85	77.72±16.86	-

memory, the two modality-specific features are downsampled with 1/8 with maxpooling and then flattened.

M9 M5+Mo+L^{S,D_p}: The extension of M5. The model is designed by adding an additional prediction discriminator D_p, which discriminates the prediction from the PET image or the CT image in the M5.

M10 M5+Mo+L_{GAN}^{S,D_M}: Our final network is the extension of M5. Modality-specific feature alignment is based on the modality discriminator D_M.

Fig. 3 shows a case of the lung tumor segmentation in PET-CT images. The blue curves are ground truth of the CT image, the green curves are ground truth of the PET image, and the red curves are the segmentation results. As can be seen in Fig. 3(b) and (n), the lung tumor contours in the two modality images are different from each other since they are not perfectly corresponded regard to its position, shape and intensities. Precision, Recall, IoU, Dice are 63.56%, 81.53%, 54.19%, 68.69% for lung tumor annotation between in PET images and CT images. The result shows that the modality-specific lung tumor segmentation is important to accurately segment lung tumor in PET images and CT images, because of the large inconsistency between these two modality images.

Our baseline consists of two segmentation blocks for the PET image and the CT image, and therefore, the corresponding lung tumor segmentation can be simultaneously obtained. We first trained the netwoork as the configuration of M1 in Table I, Precision, Recall, IoU, Dice reach 47.11%, 73.81%, 39.17%,51.74% for lung tumor segmentation in CT image, and these indices reach 54.82%, 82.5%, 47.42%, 60.02% for lung tumor segmentation in PET image. With addtional loss function L_{bce}^S to M1, Precision, IoU, Dice are improved by 2.44%, 2.05%, 2.03% for lung tumor segmentation in CT image, and these indices are improved by 6.27%, 4.61%, 4.19% for lung tumor segmentation in PET image. With addtional loss function L_{dice}^S to M2, Precision, IoU, Dice are improved by 9.15%, 4.9%, 5.12% for lung tumor segmentation in CT image, and these indices are decreased slightly for lung tumor segmentation in PET image. As all the normal thorax PET-CT slice pairs are used to train the baseline, Precision, IoU, Dice are further improved by 22.48%, 14.26%, 22.79% for lung tumor segmentation in CT image, and these indices are improved by 22.73%, 11.44%, 10.14% for lung tumor segmentation in PET image.

With the configuration of M4, Precision, Recall, IoU, Dice reach 81.28%, 70.54%, 60.37%,71.68% for lung tumor segmentation in CT image, and these indices reach 83.76%, 74.51%, 62.36%, 73.96% for lung tumor segmentation in PET image. When the two modality-specific decoder brancher DE_{CT} and DE_{PET} are integrated into the network. Compared to M4, IoU and Dice slightly decrease by 0.88% and 0.15% in CT image segmentation, and they also slightly decrease by 0.24% and 0.23% in PET image segmentation. Modality-specific feature might slightly reduce the fusion and complementarity. Compared to M4, by using MoSNet, Recall, IoU, Dice increase by 1.06%, 4.09%, 3.39%, 3.84% for lung tumor segmentation in CT image, and Recall, IoU, Dice increase by 5.12%, 3.9%, 3.76% for lung tumor segmentation in PET image. The two p values are smaller than 0.05, which shows that the modality discriminator D_M has statistically improved performance of lung tumor segmentation in PET-CT images.

Compared to MoSNet, Precision, IoU, Dice decrease by 29.11%, 18.99%, 18.55% for lung tumor segmentation in CT image, and Precision, IoU, Dice decrease by 27.78%, 19.11%,

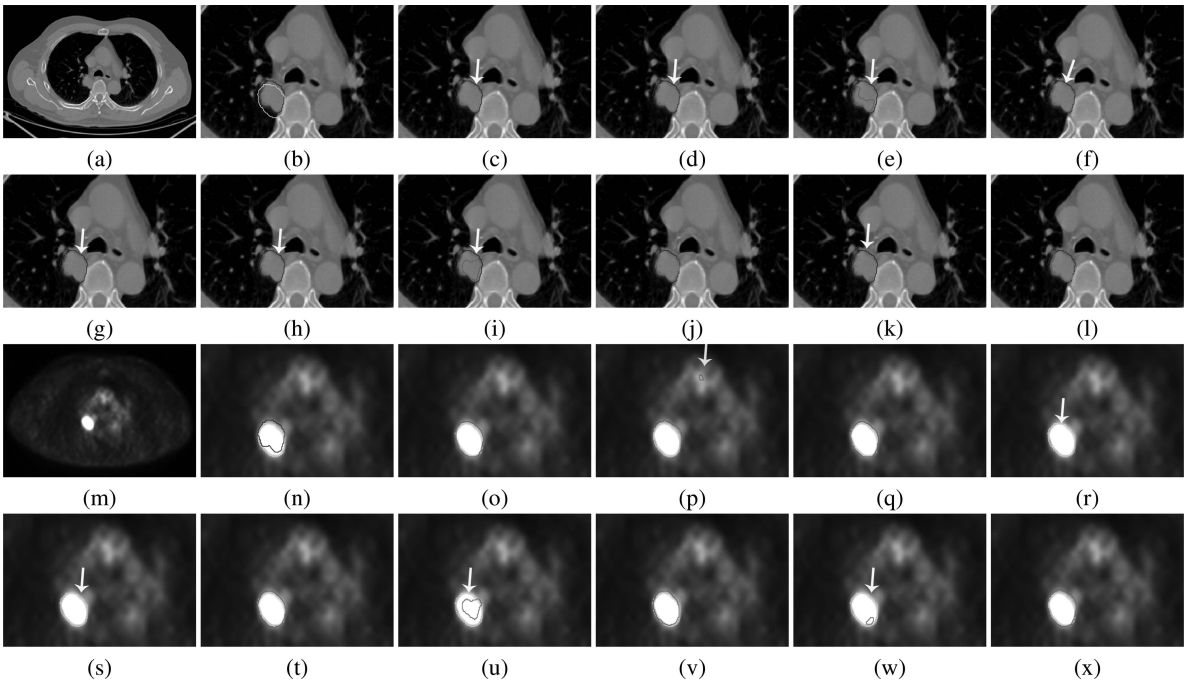

Fig. 3. Lung tumor segmentation results of different methods. The blue curves are ground truth of the CT image, the green curves are ground truth of the PET image, and the red curves are the segmentation results. The yellow arrows point out the under-segmentation and the cyan arrows point out the over-segmentation. (c)-(l) Lung tumor segmentation results of different methods in the CT image, (o)-(x) Lung tumor segmentation results of different methods in the PET image. (a) CT image, (b) CT image with ground truth of the CT image and PET image, (c) RCNet [34], (d) UNet [8], (e) nnUNet [56], (f) VNet3d [7], (g) WNet [18], (h) UNet3d [19], (i) UNet2.5d [8], (j) MSAM [9], (k) Co-learning [6], (l) MoSNet; (m) the corresponding PET image of (a), (n) PET image with ground truth of the CT image and PET image, (o) RCNet [34], (p) UNet [8], (q) nnUNet [56], (r) VNet3d [7], (s) WNet [18], (t) UNet3d [19], (u) UNet2.5d [8], (v) MSAM [9], (w) Co-learning [6], (x) MoSNet. Except (a) and (m), the rest images are locally enlarged.

18.27% for lung tumor segmentation in PET image by using M6. This shows that MMD of modality-specific features leads to negative feature alignment. When MMD is used to discriminate PET and CT prediction by using M7, Precision, Recall, IoU, Dice decrease by 4.32%, 10.46%, 8.77%, 9.21% for lung tumor segmentation in CT image, and Precision, Recall, IoU, Dice decrease by 1.7%, 13.02%, 10.33%, 10.82% for lung tumor segmentation in PET image. This shows that MMD of modality-specific prediction leads to negative prediction alignment.

Compared to MoSNet, Precision, Recall, IoU, Dice decrease by 5.69%, 2.13%, 4.79%, 4.51% for lung tumor segmentation in CT image, and Precision, IoU, Dice decrease by 1.05%, 2.03%, 1.68%, 2.28% for lung tumor segmentation in PET image by using M8. p of CT image segmentation is smaller than 0.05 but that of PET image segmentation is larger than 0.05. This shows that GRL does not improve lung tumor segmentation in CT image. Compared to MoSNet, the prediction discriminator leads to the decrease of Recall, IoU, Dice by 7.06%, 4.25%, 4.1% for lung tumor segmentation in CT image, and decrease of Recall, IoU, Dice by 3.92%, 1.08%, 1.45% for lung tumor segmentation in PET image by using M9. Although the decrease in PET image segmentation is not statistically significant, the decrease in CT image segmentation is statistically significant.

E. Comparison Against State-of-The-Art Networks

MoSNet is also compared to state-of-the-art lung tumor segmentation networks: UNet [8], RCNet [34], UNet3d [19], UNet2.5d [8], VNet3d [7], nnU-Net [56], WNet [18], Co-learning [6] and MSAM [9]. Quantitative comparisons are shown in Table II. The model training is configured as,

- C1 UNet: For the single modality image segmentation network, the CT image is fed to train the network with the corresponding manual annotation of CT images, while the PET image is fed to train the network with the corresponding manual annotation of PET images.
- C2 RCNet: The network is also trained as UNet.
- C3 nnUNet: 3D fullres model is used and trained as UNet.
- C4 VNet3d: For the single modality image 3D segmentation network, the 3D CT image patch is fed to train the network with the corresponding manual annotation of CT images, while the 3D PET image patch is fed to train the network with the corresponding manual annotation of PET images. The image patch size is $512 \times 512 \times 8$, and the moving stride is (512,512,2).
- C5 UNet2.5 d: Another segmentation layer is added to UNet. The input is the concatenation of the PET-CT image pair. The corresponding manual annotations of PET-CT images are used to compute the loss of the network.

TABLE II
COMPARISONS OF SEGMENTATION RESULTS BETWEEN MOSNET AND STATE-OF-THE-ART NETWORKS (MEAN ± STANDARD DEVIATION)

CT	Precision(%)	Recall(%)	IoU(%)	**Dice**(%)	p
UNet [8]	73.23±29.53	57.88±30.87	48.08±28.66	59.2±30.03	1.43×10^{-11}
RCNet [34]	69.25±32.17	56.56±33.17	46.47±30.46	56.79±32.45	1.55×10^{-11}
nnUNet [56]	81.77±30.26	56.17±33.79	51.15±32.86	60.12±35.18	1.31×10^{-7}
VNet3d [7]	80.67±24.84	59.63±29.75	51.84±27.78	63.14±28.5	3.57×10^{-7}
UNet2.5d [8]	62.51±28.82	75.91±22.31	50.99±24.76	63.22±26.72	1.97×10^{-7}
WNet [18]	74.84±28.62	63.19±29.95	52.63±28.71	63.48±29.51	4.97×10^{-7}
UNet3d [19]	81.6±24.51	59.36±30.04	52.78±28.13	63.93±28.44	2.67×10^{-6}
MSAM [9]	70.62±26.58	**76.37±23.12**	57±24.57	68.9±24.1	1.20×10^{-4}
Co-learning [6]	78.41±26.42	70.3±26.67	59.1±26.51	69.97±26.31	5.61×10^{-3}
Our method	**82.34±19.33**	74.63±21.84	**63.76±20.99**	**75.52±18.65**	-
PET	Precision(%)	Recall(%)	IoU(%)	**Dice**(%)	p
UNet [8]	53.52±33.67	75.69±26.04	42.15±28.32	53.39±30.09	2.40×10^{-15}
RCNet [34]	65.64±34.8	59.15±31.88	46.68±29.93	57.31±31.53	5.21×10^{-12}
UNet3d [19]	70.91±31.82	66.57±28.84	51.17±28.25	62.36±29.01	6.07×10^{-11}
UNet2.5d [8]	72.38±31.21	69.77±21.04	53.74±24.77	65.82±25.83	2.86×10^{-6}
VNet3d [7]	74.08±27.54	69.48±26.67	56.26±25	68.07±24.98	1.28×10^{-5}
nnUNet [56]	71.37±30.01	78.89±24.31	59.17±27.08	69.96±26.16	2.12×10^{-4}
MSAM [9]	79.89±25.28	72.05±23.25	59.93±24.34	71.4±23.72	5.72×10^{-4}
Co-learning [6]	81.77±24.97	71.72±26.07	60.25±25	71.54±23.82	1.36×10^{-3}
WNet [18]	78.28±22.6	75.67±23.8	60.73±22.37	72.64±21.4	1.94×10^{-3}
Our method	**82.55±20.9**	**79.63±18.85**	**66.26±19.85**	**77.72±16.86**	-

C6 WNet: For the two modality image 3D segmentation network, the 3D CT image patch and the 3D PET image patch are fed to train the network with the corresponding manual annotations. The image patch size is $512 \times 512 \times 8$, and the moving stride is (512,512,2).

C7 UNet3d: The network is trained as VNet3d.

C8 MSAM: When the network is trained with the corresponding manual annotation for the segmentation of the PET image, the CT image is fed to MSAM for the computation of the spatial attention map. When the network is trained with the corresponding manual annotation for the segmentation of the CT image, the PET image is fed to MSAM for the computation of the spatial attention map.

C9 Co-learning: Another segmentation layer is added to the original Co-learning [6]. The paired PET-CT images are respectively fed to the two encoders of Co-learning. The corresponding manual annotations of PET-CT images are used to compute the loss of the network.

Fig. 3(c)-(i) and (k) show that under-segmentation tends to be leaded for lung tumor segmentation in the CT image. RCNet, UNet, VNet3d, WNet and UNet3d do not segment the lung tumor. nnUNet, UNet2.5 d and Co-learning can segment a small part of the lung tumor. With the constraint of the PET image, MSAM can segment the lung tumor. MoSNet can accurately segment the lung tumor with modality-specific feature alignment. As can be seen in Fig. 3 (p), the lung tumor is over-segmented in the PET image by using UNet. As can be seen in Fig. 3 (r) and (s), VNet3d and WNet can not segment the lung tumor in the PET image. Fig. 3(w) shows that Co-learning under-segments the lung tumor in the PET image. MoSNet can accurately segment the lung tumor.

A comparison of the improvement on tumor segmentation performance is presented in Table II. One case of lung tumor segmentation is show in Fig. 3. As can be seen in Table II, lung tumor segmentation in the CT or PET image by using RCNet, UNet, nnUNet, UNet3d and VNet3d is inferior to those methods with PET-CT images. Compared to UNet, the combination of PET-CT images can improve the accuracy of modality-specific segmentation of the lung tumor. Recall, IoU, Dice increase by 18.03%, 2.91%, 3.94% for lung tumor segmentation in CT image, and Precision, IoU, Dice by 18.86%, 11.59%, 12.43% increase by for lung tumor segmentation in PET image by using UNet2.5 d. Although WNet used two VNet3D to respectively extract PET-CT feature, a feature fusion module was subsequently used to extract features from PET-CT feature maps [18]. The feature fusion is too late to segment lung tumor segmentation in the CT image, since the constrast of the CT image is much lower than that of the PET image. Co-learning [6] outperforms MSAM [9] for lung tumor segmentation both in the CT and PET image. Co-learning extracts and fuses PET-CT-specific feature while MSAM only fused a spatial attention map to the other modality image. MoSNet consists of two additional modality-specific decoder branches, and therefore, it outperforms state-of-the-art networks. The modality discriminator is further used to softly reduce the modality-specific feature discrepancy, and allows the network to preserve complementarity between PET and CT images.

F. Discussion

In this paper, we present a novel deep neural network framework to automatically segment the lung tumor in PET-CT images. The results presented in the previous sections have shown our model's effectiveness and advancement in the multiple modality image segmentation.

With the cGAN framework, a novel model is designed with a dual stream encoder to extract diverse and complementary

features of PET images and CT images, a decoder to fuse these different multiple modality features and two decoders to preserve modality-specific features of PET images and CT images. As shown in the ablation experiments, M4 demonstrates that lung tumor segmentation is beneficial from the diverse and complementary features of PET images and CT images. Although the modality-specific features of PET images and CT images slightly reduce the accuracy of lung tumor segmentation, the adversarial method can minimize an approximate modality discrepancy through an adversarial objective with respect to a modality discriminator and reserve modality-common representation. After the modality discriminator D_M is added to M5, MoSNet largely improve the lung tumor segmentation both in PET images and CT images, compared to M4. Precision. This adversarial method improves the representation power of the network for modality-specific lung tumor segmentation in PET images and CT images.

With regard to modality-specific feature alignment, several adversarial methods have been tested in the ablation experiments. MMD is used to discriminate modality-specific features and predictions of PET and CT images. MMD leads to negative modality-specific features and prediction alignment. In addition, the complexity of MMD in the reproducing kernel Hilbert space is large to compute for both modality-specific features and prediction alignment. GRL can also be used to minimize an approximate modality discrepancy; however, the fully connected layers in GRL need to consume huge memory for modality-specific features. Therefore, the two modality-specific features are down-sampled and flattened, which reduces the performance of the discriminator. On contrary, the modality discriminator D_M consists of convolutional layers and is flexible to the multiple channel features with the original size.

The other main contributions of this paper are modality-specific lung tumor segmentation and usage of normal thorax PET-CT slice pairs. As shown in the ablation experiments and Fig. 3, the lung tumor contours in the two modality images are different from each other due to the modality discrepancy and patient movement, which leads to different ground truth of the two modality images. Therefore, it is necessary to accomplish the modality-specific lung tumor segmentation of PET-CT images. In addition, many previous researches only trained and tested their models with the 2D thorax slices containing tumors [6], [9], [34]. In contrast, our models are trained with normal thorax PET-CT slice pairs. The experiments show that the normal thorax PET-CT slice pairs can largely improve the segmentation of the lung tumors since there are several neighboring organs, e.g. liver, heart, and muscles, and they share the similar intensities both in CT images and PET images between the tumor and its neighboring organs.

V. Conclusion and Future Work

In this paper, a modality-specific segmentation network is proposed for lung tumor segmentation in PET-CT images. Due to respiration and movment, the lung tumor vary largely in PET images and CT images. Considering the complementarity and inconsistency between PET images and CT images, our method can simultaneously obtain the corresponding lung tumor segmentations of PET images and CT images. MoSNet can learn a modality-fused representation to encode the common feature of lung tumor in PET images and CT images, and meanwhile, an adversarial method is also proposed to minimize an approximate modality discrepancy through an adversarial objective with respect to a modality discriminator and preserve modality-common representation. Therefore, MoSNet improves the performance of the network for modality-specific lung tumor segmentation in PET images and CT images.

One limitation is that our model is designed for 2D thorax PET-CT slices. As shown in the comparison experiments, the 2.5D and 3D UNets outperform the 2D UNet. Future work will aim to utilize the neighboring relationships between connected 2D thorax slices.

References

[1] J. Ferlay et al., "Cancer statistics for the year 2020: An overview," *Int. J. Cancer*, vol. 149, no. 4, pp. 778–789, 2021.

[2] S. McGuire, "World Cancer report 2014. Geneva, Switzerland : World Health Organization, international agency for research on Cancer," WHO Press, 2015, *Adv. Nutrition*, vol. 7, pp. 418–419, 2016.

[3] Y. E. Erdi et al., "Radiotherapy treatment planning for patients with non-small cell lung Cancer using positron emission tomography (PET)," *Radiotherapy Oncol.*, vol. 62, no. 1, pp. 51–60, 2002.

[4] C. Greco, K. Rosenzweig, G. L. Cascini, and O. Tamburrini, "Current status of PET/CT for tumour volume definition in radiotherapy treatment planning for non-small cell lung Cancer (NSCLC)," *Lung Cancer*, vol. 57, no. 2, pp. 125–134, 2007.

[5] Y.-S. Kao and J. Yang, "Deep learning-based auto-segmentation of lung tumor PET/CT scans: A systematic review," *Clin. Transl. Imag.*, vol. 10, pp. 217–223, 2022.

[6] A. Kumar, M. Fulham, D. Feng, and J. Kim, "Co-learning feature fusion maps from PET-CT images of lung cancer," *IEEE Trans. Med. Imag.*, vol. 39, no. 1, pp. 204–217, Jan. 2020.

[7] L. Li, X. Zhao, W. Lu, and S. Tan, "Deep learning for variational multimodality tumor segmentation in PET/CT," *Neurocomputing*, vol. 392, pp. 277–295, 2020.

[8] Y. Lu, J. Lin, S. Chen, H. He, and Y. Cai, "Automatic tumor segmentation by means of deep convolutional U-Net with pre-trained encoder in PET images," *IEEE Access*, vol. 8, pp. 113636–113648, 2020.

[9] X. Fu, L. Bi, A. Kumar, M. Fulham, and J. Kim, "Multimodal spatial attention module for targeting multimodal PET-CT lung tumor segmentation," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 9, pp. 3507–3516, Sep. 2021.

[10] H. Ligtenberg et al., "Modality-specific target definition for laryngeal and hypopharyngeal cancer on FDG-PET, CT and MRI," *Radiother. Oncol.*, vol. 123, no. 1, pp. 63–70, 2017.

[11] G. W. Goerres, E. Kamel, T.-N. H. Heidelberg, M. R. Schwitter, C. Burger, and G. K. von Schulthess, "PET-CT image co-registration in the thorax: Influence of respiration," *Eur. J. Nucl. Med. Mol. Imag.*, vol. 29, no. 3, pp. 351–360, 2002.

[12] S. J. Rosenbaum, T. Lind, G. Antoch, and A. Bockisch, "False-positive FDG PET uptake- The role of PET/CT," *Eur. Radiol.*, vol. 16, no. 5, pp. 1054–1065, 2006.

[13] W. Ju, D. Xiang, B. Zhang, L. Wang, I. Kopriva, and X. Chen, "Random walk and graph cut for co-segmentation of lung tumor on PET-CT images," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5854–5867, Dec. 2015.

[14] S. S. Shim et al., "Non-small cell lung cancer: Prospective comparison of integrated FDG PET/CT and CT alone for preoperative staging," *Radiology*, vol. 236, no. 3, pp. 1011–1019, 2005.

[15] G. Antoch et al., "Non-small cell lung cancer: Dual-modality PET/CT in preoperative staging," *Radiology*, vol. 229, no. 2, pp. 526–533, 2003.

[16] Q. Song et al., "Optimal co-segmentation of tumor in PET-CT images with context information," *IEEE Trans. Med. Imag.*, vol. 32, no. 9, pp. 1685–1697, Sep. 2013.

[17] U. Bağci, J. Yao, J. Caban, E. Turkbey, O. Aras, and D. J. Mollura, "A graph-theoretic approach for segmentation of PET images," in *Proc. IEEE Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2011, pp. 8479–8482.

[18] L. Xu et al., "Automated whole-body bone lesion detection for multiple myeloma on 68Ga-Pentixafor PET/CT imaging using deep learning methods," *Contrast Media Mol. Imag.*, vol. 2018, 2018, Art. no. 2391925.
[19] Z. Zhong et al., "3D fully convolutional networks for co-segmentation of tumors on PET-CT images," in *Proc. IEEE 15th Int. Symp. Biomed. Imag.*, 2018, pp. 228–231.
[20] G. Chen et al., "Automatic pathological lung segmentation in low-dose CT image using eigenspace sparse shape composition," *IEEE Trans. Med. Imag.*, vol. 38, no. 7, pp. 1736–1749, Jul. 2019.
[21] X. Zhao, L. Li, W. Lu, and S. Tan, "Tumor co-segmentation in PET/CT using multi-modality fully convolutional neural network," *Phys. Med. Biol.*, vol. 64, no. 1, 2018, Art. no. 015011.
[22] Y. E. Erdi et al., "Segmentation of lung lesion volume by adaptive positron emission tomography image thresholding," *Cancer: Interdiscipl. Int. J. Amer. Cancer Soc.*, vol. 80, no. S12, pp. 2505–2509, 1997.
[23] W. Jentzen, L. Freudenberg, E. G. Eising, M. Heinze, W. Brandau, and A. Bockisch, "Segmentation of PET volumes by iterative image thresholding," *J. Nucl. Med.*, vol. 48, no. 1, pp. 108–114, 2007.
[24] S. Nehmeh et al., "An iterative technique to segment pet lesions using a Monte Carlo based mathematical model," *Med. Phys.*, vol. 36, no. 10, pp. 4803–4809, 2009.
[25] X. Geets, J. A. Lee, A. Bol, M. Lonneux, and V. Grégoire, "A gradient-based method for segmenting FDG-PET images: Methodology and validation," *Eur. J. Nucl. Med. Mol. Imag.*, vol. 34, no. 9, pp. 1427–1438, 2007.
[26] S. Belhassen and H. Zaidi, "A novel fuzzy C-means algorithm for unsupervised heterogeneous tumor quantification in PET," *Med. Phys.*, vol. 37, no. 3, pp. 1309–1324, 2010.
[27] C. Ballangan, X. Wang, M. Fulham, S. Eberl, and D. D. Feng, "Lung tumor segmentation in PET images using graph cuts," *Comput. Methods Programs Biomed.*, vol. 109, no. 3, pp. 260–268, 2013.
[28] I. Jafar, H. Ying, A. F. Shields, and O. Muzik, "Computerized detection of lung tumors in PET/CT images," in *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2006, pp. 2320–2323.
[29] Y. Guo et al., "Automatic lung tumor segmentation on PET/CT images using fuzzy Markov random field model," *Comput. Math. Methods Med.*, vol. 2014, 2014, Art. no. 401201.
[30] J. Soltani-Nabipour, A. Khorshidi, and B. Noorian, "Lung tumor segmentation using improved region growing algorithm," *Nucl. Eng. Technol.*, vol. 52, no. 10, pp. 2313–2319, 2020.
[31] S. Vijh, R. Sarma, and S. Kumar, "Lung tumor segmentation using marker-controlled watershed and support vector machine," *Int. J. E-Health Med. Commun.*, vol. 12, no. 2, pp. 51–64, 2021.
[32] D. Han et al., "Globally optimal tumor segmentation in PET-CT images: A graph-based co-segmentation method," in *Proc. Biennial Int. Conf. Inf. Process. Med. Imag.*, 2011, pp. 245–256.
[33] U. Bağci et al., "Joint segmentation of anatomical and functional images: Applications in quantification of lesions from PET, PET-CT, MRI-PET, and MRI-PET-CT images," *Med. Image Anal.*, vol. 17, no. 8, pp. 929–945, 2013.
[34] J. Jiang et al., "Multiple resolution residually connected feature streams for automatic lung tumor segmentation from CT images," *IEEE Trans. Med. Imag.*, vol. 38, no. 1, pp. 134–144, Jan. 2019.
[35] H. Hu, Q. Li, Y. Zhao, and Y. Zhang, "Parallel deep learning algorithms with hybrid attention mechanism for image segmentation of lung tumors," *IEEE Trans. Ind. Inform.*, vol. 17, no. 4, pp. 2880–2889, Apr. 2021.
[36] S. Pang et al., "CTumorGAN: A unified framework for automatic computed tomography tumor segmentation," *Eur. J. Nucl. Med. Mol. Imag.*, vol. 47, no. 10, pp. 2248–2268, 2020.
[37] W. Gan et al., "Automatic segmentation of lung tumors on CT images based on a 2D & 3D hybrid convolutional neural network," *Brit. J. Radiol.*, vol. 94, 2021, Art. no. 20210038.
[38] S. Tyagi, S. N. Talbar, and A. Mahajan, "A novel approach of lung tumor segmentation using a 3D deep convolutional neural network," in *Handbook of Research on Applied Intelligence for Health and Clinical Informatics*. Hershey, PA, USA: IGI Global, 2022, pp. 1–16.
[39] J. Li, H. Chen, Y. Li, Y. Peng, J. Sun, and P. Pan, "Cross-modality synthesis aiding lung tumor segmentation on multi-modal MRI images," *Biomed. Signal Process. Control*, vol. 76, 2022, Art. no. 103655.
[40] P. Dutande, U. Baid, and S. Talbar, "Deep residual separable convolutional neural network for lung tumor segmentation," *Comput. Biol. Med.*, vol. 141, 2022, Art. no. 105161.
[41] J. Zhao, M. Dang, Z. Chen, and L. Wan, "DSU-Net: Distraction-sensitive U-Net for 3D lung tumor segmentation," *Eng. Appl. Artif. Intell.*, vol. 109, 2022, Art. no. 104649.
[42] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7167–7176.
[43] S. Li et al., "Domain conditioned adaptation network," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11386–11393.
[44] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," 2014, *arXiv:1412.3474*.
[45] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. 32nd. Int. Conf. Mach. Learn.*, 2015, pp. 97–105.
[46] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 2058–2065.
[47] B. Sun and K. Saenko, "Deep Coral: Correlation alignment for deep domain adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 443–450.
[48] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 14 1–9.
[49] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," 2016, *arXiv:1611.02200*.
[50] Y. Ganin et al., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, 2016.
[51] M. L. Junguang Jiang and B. Fu, "Transfer-learning-library," 2020. [Online]. Available: https://github.com/thuml/Transfer-Learning-Library
[52] F. Mahmood et al., "Deep adversarial training for multi-organ nuclei segmentation in histopathology images," *IEEE Trans. Med. Imag.*, vol. 39, no. 11, pp. 3257–3267, Nov. 2020.
[53] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada, "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3723–3732.
[54] Y. Zhang, T. Liu, M. Long, and M. Jordan, "Bridging theory and algorithm for domain adaptation," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7404–7413.
[55] P. A. Yushkevich et al., "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability," *Neuroimage*, vol. 31, no. 3, pp. 1116–1128, 2006.
[56] F. Isensee, P. F. Jaeger, S. 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, 2021.