Computerized Medical Imaging and Graphics 88 (2021) 101851

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

DiSegNet: A deep dilated convolutional encoder-decoder architecture for lymph node segmentation on PET/CT images

Guoping Xu [a,b,c], Hanqiang Cao [b], Jayaram K. Udupa [c,*], Yubing Tong [c], Drew A. Torigian [c]

[a] School of Computer Sciences and Engineering, Wuhan Institute of Technology, Wuhan, Hubei, 430205, China
[b] School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, Hubei, 430074, China
[c] Medical Image Processing Group, 602 Goddard Building, 3710 Hamilton Walk, Department of Radiology, University of Pennsylvania, Philadelphia, PA, 19104, United States

ARTICLE INFO

Keywords:
Convolutional neural network
Lymph node segmentation
Positron emission tomography/computed tomography (PET/CT)
Dilated convolution
Imbalance class

ABSTRACT

Purpose: Automated lymph node (LN) recognition and segmentation from cross-sectional medical images is an important step for the automated diagnostic assessment of patients with cancer. Yet, it is still a difficult task owing to the low contrast of LNs and surrounding soft tissues as well as due to the variation in nodal size and shape. In this paper, we present a novel LN segmentation method based on a newly designed neural network for positron emission tomography/computed tomography (PET/CT) images.

Methods: This work communicates two problems involved in LN segmentation task. Firstly, an efficient loss function named cosine-sine (CS) is proposed for the voxel class imbalance problem in the convolution network training process. Second, a multi-stage and multi-scale Atrous (Dilated) spatial pyramid pooling sub-module, named MS-ASPP, is introduced to the encoder-decoder architecture (SegNet), which aims to make use of multi-scale information to improve the performance of LN segmentation. The new architecture is named DiSegNet (Dilated SegNet).

Results: Four-fold cross-validation is performed on 63 PET/CT data sets. In each experiment, 10 data sets are selected randomly for testing and the other 53 for training. The results show that we reach an average 77 % Dice similarity coefficient score with CS loss function by trained DiSegNet, compared to a baseline method SegNet by cross-entropy (CE) with 71 % Dice similarity coefficient.

Conclusions: The performance of the proposed DiSegNet with CS loss function suggests its potential clinical value for disease quantification.

1. Introduction

Lymph node (LN) segmentation aims to assign a categorical label to every voxel in an image, which plays an important role in medical image analysis and disease quantification. Yet, manual detection and measurement/segmentation of LNs in images by human observers is time-consuming and error prone (Feulner et al., 2013). Moreover, LNs are difficult to recognize and segment owing to the low contrast between LNs and surrounding soft tissues as well as due to the variation in nodal size and shape.

This topic has received much attention in recent decades. Discriminative learning and a spatial prior probability have been used for LN detection and segmentation in chest computed tomography (CT) images (Feulner et al., 2013). The discriminative model was used to detect LNs from their appearance combined with the anatomical prior knowledge, and the graph cut method was used to segment LNs. In (Barbu et al., 2012a), the authors proposed a learning-based approach that used marginal space learning for LN segmentation in the axillary region. Pathological LNs were detected and segmented in (Hoogi et al., 2017). The method employed machine learning techniques such as marginal space learning, convolutional neural network, and active contour models for organ detection, LN detection, and LN segmentation. All of the methods above are based on the hand-crafted features or prior knowledge to recognize and delineate LNs.

Recently, fully convolutional networks (FCNs) (Long et al., 2015) were used for pixel-wise semantic segmentation tasks due to the development of deep learning methods, especially given the success of deep convolutional neural network (DCNN) models such as AlexNet

* Corresponding author.
E-mail address: jay@pennmedicine.upenn.edu (J.K. Udupa).

https://doi.org/10.1016/j.compmedimag.2020.101851
Received 28 June 2020; Received in revised form 20 November 2020; Accepted 15 December 2020
Available online 29 December 2020

(Krizhevsky et al., 2012), VGG (Simonyan and Zisserman, 2014), and ResNet(He et al., 2016). Many semantic segmentation architectures after FCNs appeared used the encoder and decoder structure, such as SegNet (Badrinarayanan et al., 2017) and U-Net (Ronneberger et al., 2015). However, there still exists a problem in the decoder part by recovering the low-resolution feature maps owing to the max-pooling operation. In (Chen et al., 2018), the dilated (or Atrous) convolution was proposed to overcome this problem, where Atrous-spatial-pyramid-pooling (ASPP) with various dilation rates was proposed to robustly segment objects at multiple scales. The DeepLab architecture has been designed by the integration of a ASPP sub-module in order to capture objects and context at multiple scales. Yet, it only uses one ASPP sub-module after the fifth max-pooling layer and did not exploit ASPP in different stages after the max-pooling operation that may miss some important features, especially from the small size of LNs. Moreover, it did not discuss the order that was employed for ASPP in the architecture, which is still an open question in terms of which layer/-layers should use ASPP in the architecture of DCNN.

Although much success has been achieved in semantic segmentation, there are only a few works that employ semantic segmentation neural network for LN segmentation. In (Bouget et al., 2019), the authors combined U-Net and Mask R-CNN for segmentation and detection of mediastinal LNs. However, this method needs to train two DCNNs, which increased many repeat computations for feature learning because feature maps from U-Net can also be used in Mask R-CNN (Ren et al., 2017). In (Oda, 2018), the 3D U-Net was trained to segment LNs and other anatomical structures on contrast-enhanced chest CT volumes. Yet, it incurs much computation cost and memory consumption, especially GPU memory, given the 3D convolution operation.

There are two issues which are not addressed by the LN segmentation methods reviewed above. Firstly, none of the proposed methods considered the imbalance in voxel classes between the LNs and the remaining part of the volumes, which will impede the training efficiency. In (Lin et al., 2017), they proposed a focal loss function for the dense object detection problem, which focused on the imbalance of the bounding box between the objects and the background by down-weighting the loss from the well-classified examples. Inspired by the focal loss function, an exponential loss function (Xu et al., 2020) is proposed for the slice classification task, which aims to classify each slice as to whether or not it contains pathological LNs on PET/CT. However, the problem of imbalance is more severe in the LN segmentation task owing to the fact that most voxels do not constitute LNs. Second, the multi-scale information from feature maps is helpful to do semantic segmentation in natural images (Chen et al., 2018), which did not used in LN segmentation.

In this work, there are three novelties:

Firstly, considering the merit of multi-scale feature analysis by ASPP and training time and efficiency, a new strategy is proposed by using more than one ASPP sub-module called multi-stage Atrous-spatial-pyramid-pooling (MS-ASPP) after the max-pooling layer in the encoder part, which could provide more context information into the decoder to aid with LN segmentation. We named this DiSegNet, which involves the integration of MA-ASPP into the SegNet architecture.

Second, we tested the loss functions that are used to train the semantic segmentation network such as cross-entropy and Dice loss (Sudre et al., 2017), and found that these loss functions did not consider the imbalance of voxel classes. In (Lin et al., 2017), the authors designed a focal loss function for the imbalance problem of dense object detection. A new exponential loss function is proposed in (Xu et al., 2020) for pathological LN classification, which aims to handle the imbalance of slice classes that include and do not include LNs. Inspired by the idea of focal loss function (FL) and exponential loss function (EL), we propose a novel loss function named cosine-sine (CS) loss function to deal with the imbalance of voxel classes for LN segmentation in this study. Compared to FL and EL, the proposed CS loss function will up-weight the loss from misclassified voxels while down-weighting the loss from well-classified voxels, which facilitates the pathological LN segmentation task.

Third, to our best knowledge, this is the first report of pathological LN segmentation in positron emission tomography/computed tomography (PET/CT) images of thorax based on DCNN. We also compared our method to other published papers that perform LN segmentation on CT or PET/CT volumes.

In this paper, the definitions of the cross-entropy loss function and focal loss function are first introduced. Then, the materials are introduced in Section 2.1 and a novel cosine-sine (CS) loss function is proposed in Section 2.2.1. Two variants of the DiSegNet architecture that employ the MS-ASPP are introduced in section 2.3.2. In part 3 and part 4, the experiments and discussion will be shown. Finally, the conclusion is provided.

2. Materials and methods

2.1. Materials

This retrospective study was conducted following approval from the Institutional Review Board at the University of Pennsylvania (UPenn) along with a Health Insurance Portability and Accountability Act waiver. The data set was retrospectively selected from our health system patient image database by a board-certified radiologist (D.A.T), which consists of 63 ^{18}F-fluorodeoxyglucose (FDG) PET/CT image data sets from 63 subjects with either Hodgkin lymphoma or diffuse large B-cell lymphoma (DLBCL). Each CT image data set utilized in this study consisted of an average of 70 axial slices covering the entire thorax, with a mean pixel size of 1.14 mm $\times$ 1.14 mm and a mean slice spacing of 3.75 mm without intravenous contrast material. Each PET image data set utilized in this study consisted of an average of 70 axial slices covering the entire thorax, with a mean pixel size of 4 mm x 4 mm and a mean slice spacing of 4 mm. We resized all PET images to CT size by using linear interpolation in order to match images and voxels from PET and CT. Abnormal LN delineations in the thorax on the PET/CT data sets were first performed to serve as ground truth, where abnormal LNs were initially identified from PET and CT images by a board-certified radiologist (co-author D.A.T), and then LN masks were subsequently created by interactive thresholding on the PET images followed by manual adjustment using information from the accompanying CT images. Note that only the abnormal LNs in the mediastinal and hilar portion of thorax were included in our study, which has 176 LNs and 38 LNs respectively.

In this study, four-fold cross-validation was performed on 63 PET/CT data sets for all different network architectures, such as FCN, U-Net and DiSegNet. In each experiment, 10 data sets were selected randomly for testing and the other 53 for training. Moreover, we did not use a validation set for early stopping of optimization.

Pathological LN segmentation can be divided into two tasks: LN zone recognition and LN segmentation in LN zones. A strength of LN zone recognition is that it excludes much irrelevant territory that does not include LNs, thereby reducing the amount of required computation for LN segmentation. In our prior work, thoracic LN zones are recognized by using our automatic anatomy recognition (AAR) method proposed in (Udupa, 2014) and (Xu et al., 2018). In the current work, we will only focus on the second task, namely pathological LN segmentation in the recognized thoracic LN zones. For more details of AAR, please see (Udupa, 2014; Xu et al., 2018).

2.2. Methods

In this section, we will investigate different loss functions such as cross-entropy and focal loss. Then, the definition of the proposed cosine-sine (CS) loss function is introduced.

2.2.1. Definition of loss function

The cross-entropy (CE) loss for binary classification is as follows:

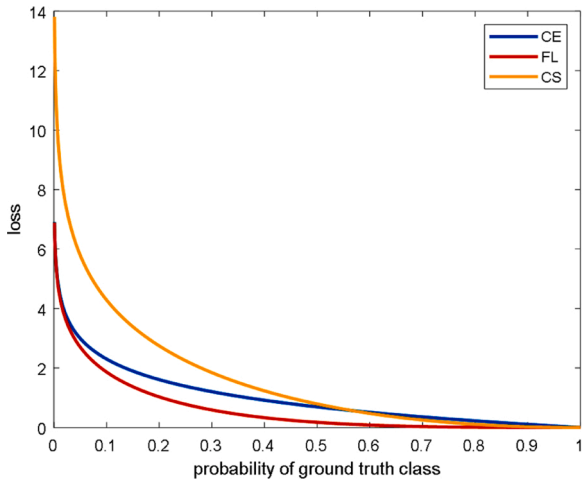


Fig. 1. Cross-entropy (CE) loss function, focal loss (FL) function with $\gamma = 2$, and cosine-sine (CS) loss function with $\gamma = 2$.

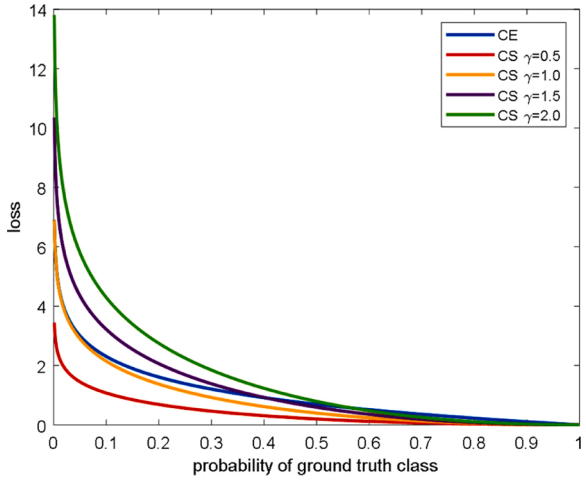


Fig. 2. The CE loss and CS loss with different modulating parametersγ. It can be seen that CS loss will reduce the relative loss for well-classified examples.

$$\mathrm{CE}(p,y) = \begin{cases} -\log(p), & if\ y = 1 \\ -\log(1-p), & \text{otherwise} \end{cases} \tag{1}$$

Here, the above $y \in \{1, 0\}$ means a class label of LNs and background, and $p \in \{1, 0\}$ is the model's estimated probability of the LN class label. For notational convenience, we define p_t:

$$p_t = \begin{cases} p & if y = 1 \\ 1-p & \text{otherwise} \end{cases} \tag{2}$$

There, we can rewrite CE(p,y) = CE(p_t) = -log(p_t). The CE loss function is shown in Fig. 1.

In (Lin et al., 2017), the authors introduced a focal loss function to deal with the imbalance of object detection problem. The definition of focal loss is as follows:

$$\mathrm{FL}(p_t) = -(1-p_t)^{\gamma}\log(p_t) \tag{3}$$

It is a variant of cross-entropy loss function that incorporates a modulating factor. It can down-weight easily classified examples and focus training on hard-to-classify examples. The FL loss function can be seen in Fig. 1 with $\gamma = 2$.

We extend the idea of the focal loss function and propose a new loss function called cosine-sine (CS) loss function in this paper. The cosine-sine (CS) cross-entropy loss function is designed to address the class imbalance (e.g., 1:2000 in one slice) problem in the number of voxels between the foreground (LNs) and background classes during the training process. We introduce the CS loss starting from the definition as follows:

$$\mathrm{CS}(p_t) = -\gamma(\cos(p_t) - \alpha\sin(p_t))\log(p_t) \tag{4}$$

The CS loss is shown in Fig. 1 with$\alpha = 0.64$ and $\gamma = 2$. Compared to the CE and FL loss functions, there are two properties of the CS loss. (1) When a voxel is misclassified and the estimated probability is small, it will increase the loss value with γ greater than 1. Therefore, the training network could focus on the misclassified examples (voxels). However, as the probability increased, the loss for these well-classified examples (voxels) will be decreased in terms of CE loss. Meanwhile, the CS loss will not reduce so quickly for the well-classified examples compared to the FL loss. This indicates that the training network is still able to reduce the loss value from well-classified examples (2) The modulating parameter γ is flexible for use in different situations. For example, if γ is less than 1, the CS loss will become the FL loss form. The CS loss can be seen for several values of γ in Fig. 2 below.

2.2.2. *Multi-stage atrous spatial pyramid pooling (MS-ASPP) and DiSegNet*

Firstly, we will introduce the concept of dilated convolution and then show how to use it in the SegNet architecture. In a one-dimensional space, the output of dilated convolution is defined as:

$$y[i] = \sum_{k=1}^{K} x[i + r{\cdot}k]w[k] \tag{5}$$

where $x[i]$ is the 1-D input signal, $y[i]$ is the output signal, and $w[k]$ represents a filter with the length K. The rate parameter r corresponds to the dilated rate. The filter will become the standard convolution with the

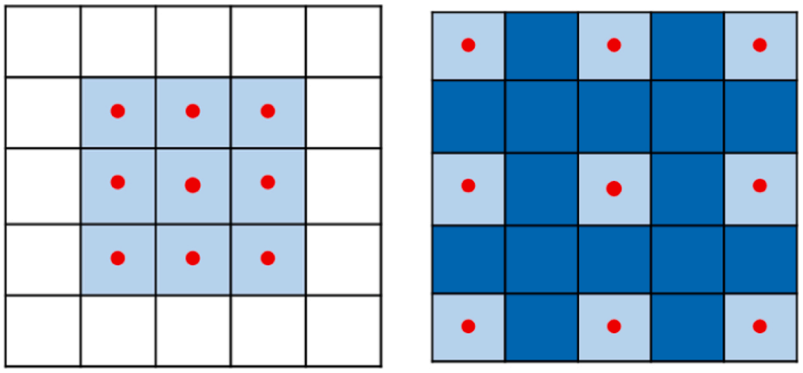


Fig. 3. Illustration of dilated convolution. Left: Standard convolution with kernel size 3 × 3; Right: Dilated convolution with the kernel size 3 × 3, a dilation rate r = 2, which will enlarge the reception filed with 5 × 5. The dark blue boxes will insert zeros when performing dilated convolution (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

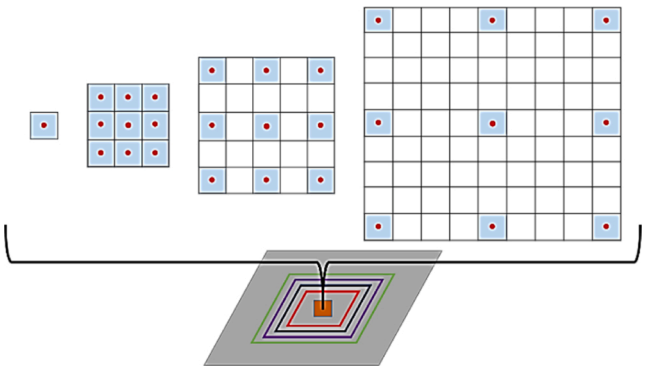


Fig. 4. The structure of ASPP. It contains one 1 × 1 convolution kernel with the rate 1, and three 3 × 3 convolution kernels with the rate 1, 2, and 4, respectively.

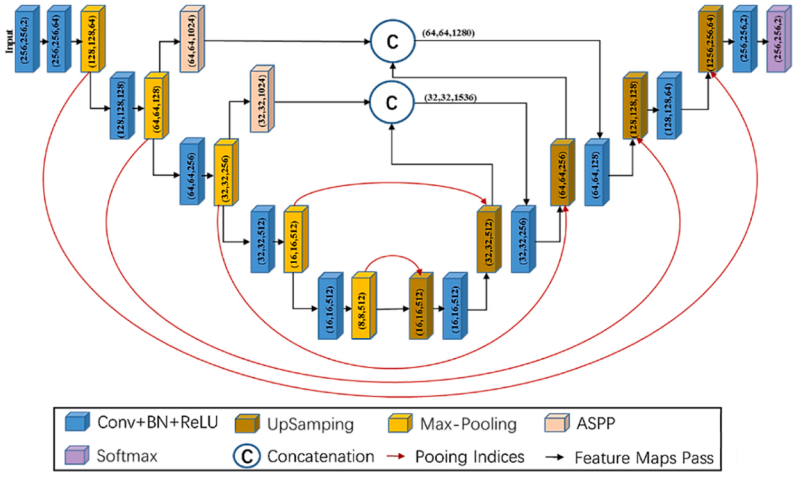
rate $r = 1$.

In the LN segmentation system, 2-D dilated convolution is performed by inserting zeros between filter values. For a convolution kernel with size $k \times k$, the size of the resulting dilated filter is k_d x× k_d, where $k_d = k + (k - 1) \times (r - 1)$. The illustration of dilated convolution is shown in Fig. 3 below with kernel size 5 × 5, the light blue color with red points indicating the kernel values and the dark blue color indicating insertions of zeros. The dilated convolution offers an elegant way to control the receptive field and maintain the high resolution of feature maps by the dilated rate of the corresponding layer, which can also obtain the result after the max-pooling operation.

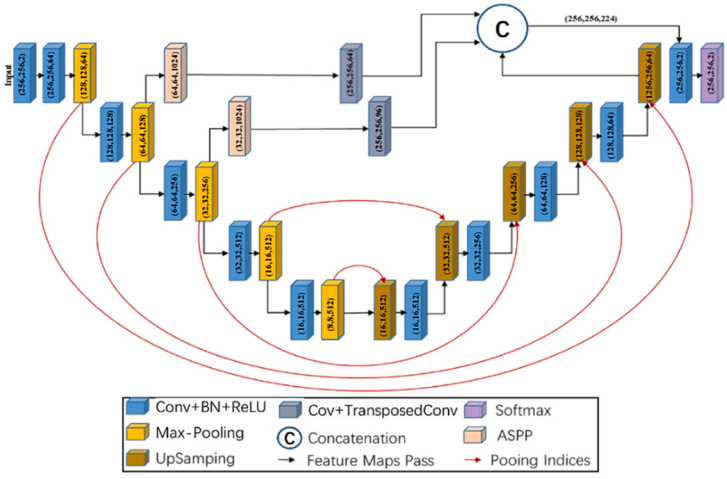
Here, we introduce a multi-stage Atrous-spatial-pyramid-pooling (MS-ASPP) sub-module that can be integrated seamlessly into the original SegNet architecture to improve the final LN segmentation performance. MS-ASPP is composed of Atrous-spatial-pyramid-pooling (ASPP) module, which can be seen in Fig. 4.

Compared to the original ASPP proposed in (Chen et al., 2018), we introduced a 1 × 1 convolution kernel from the Inception module (Szegedy et al., 2015), which includes a smaller number of more spatially spread out clusters than other larger convolution sizes. Moreover, we reduce the rate to 2 and 4 by considering the size of LNs in PET/CT.

The ASPP can also deal with the "gridding" effect from the dilated convolution as reported in (Chen et al., 2018). In our study, we use ASPP in the multi-stage after max-pooling of the SegNet framework. The ASPP can extract more local contextual information in the previous stage at a different dilation rate, and can also extract more global contextual information in the subsequent stage after max-pooling, which could help to recover LN boundaries in the decoder part. Another benefit of MS-ASPP is that it can use arbitrary dilation rates in different stages in a parallel way to train network, which will not increase training time. Moreover, the MS-ASPP can be integrated with other semantic networks



(a) DiSegNet-1: Two sub-modules combined with each upsampling layer.



(b) DiSegNet-2: Two sub-modules combined with the last upsampling layer.

Fig. 5. An illustration of two DiSegNet architectures which add multi-stage atrous spatial pyramid pooling.

like U-Net and FCN which use the same encoder and decoder structure.

In this study, we designed two kinds of architecture that use MS-ASPP in SegNet, named DiSegNet-1 and DiSegNet-2 in order to study their differences from the original SegNet. The DiSegNet is inspired by SegNet (Badrinarayanan et al., 2017) which is composed of encoder and decoder modules. The encoder part performs object recognition, (for example, LNs), by a convolutional network, which includes convolution, batch normalization, rectified-linear unit (ReLU), and max-pooling. The decoder part completes the delineation task of objects, which is composed of upsampling, convolution, batch normalization, and ReLU, where softmax is employed in the last layer. The architecture can be seen in Fig. 5 below. In DiSegNet-1, two ASPP modules are employed after max-pooling. The output feature maps from ASPP are concatenated with the feature maps after the corresponding upsampling operation in the decoder. In DiSegNet-2, the feature maps from two ASPPs are processed by 1×1 convolution and transposed convolution, and recover the size 256×256 of each feature map which is equal to the output of the last upsampling operation from the decoder. Then, the feature maps from two ASPPs and one upsampling operation are concatenated which help LN segmentation by providing more contextual information in the decoder.

In summary, the main difference of DiSegNet and SegNet lies in the multi-stage ASPP modules that are added in to the network as shown in Fig. 5. We use Atrous-spatial-pyramid-pooling (ASPP) after the max-pooling operation to extract features on two different resolution feature maps from max-pooling layer. The output feature maps of MS-ASPP combined with the feature maps from the upsampling operation are used for boundary delineation in the decoder module. Compared to U-Net which passes the feature maps of the encoder directly into the decoder, our proposed DiSegNet make use of an ASPP module which helps to extract more contextual information from the encoder.

3. Results

In this work, we utilize the VGG16 framework by discarding the fully connected layers combined with MS-ASPP module in the encoder part. The stochastic gradient descent with momentum optimizer is used. The momentum value, the initial learning rate, and the minimum batch size are 0.9, 0.001 and 4, respectively. The proposed cosine-sine (CS) loss function is employed as the objective function for training the network, and the cross-entropy loss, focal loss, and dice loss functions are used as control groups. Considering the imbalance of the class labels between LN voxels and the background, we use median frequency balancing (Eigen and Fergus, 2015), where the weight assigned to the loss function is calculated by the median of class frequencies divided by the class frequency from the training set. The augmentation strategy is also employed with random translation in the horizontal and vertical directions ranging from -10 to 10 voxels. The MSRA method described in (Kaiming et al., 2015) is employed for weight initialization. We iteratively updated models for 40 epochs, where each epoch refers to a block of iterations throughout the whole training set.

We use SegNet, FCN16 s, and DeepLabv3+ as benchmark for comparison with DiSegNet under different loss functions. Two variants of DiSegNet were designed in order to test how to layout ASPP in the SegNet. The architecture of DiSegNet-1 is shown in Fig. 5(a). The

Table 1
Comparison of different networks with various loss functions for 40 epochs. The mean and SD of Dice similarity coefficient (DSC) are displayed.

Loss		CE	DL	FL	CS
Network		DSC	DSC	DSC	DSC
SegNet	Mean	0.71	0.65	0.56	0.72
	SD	0.04	0.04	0.09	0.01
FCN16s	Mean	0.60	0.68	0.53	0.69
	SD	0.03	0.08	0.04	0.06
DeepLabv3+	Mean	0.74	0.71	0.75	0.74
	SD	0.08	0.07	0.07	0.01
DiSegNet-1	Mean	0.75	0.74	0.71	**0.77**
	SD	0.06	0.04	0.05	0.05
DiSegNet-2	Mean	0.71	0.74	0.68	0.72
	SD	0.07	0.05	0.07	0.04

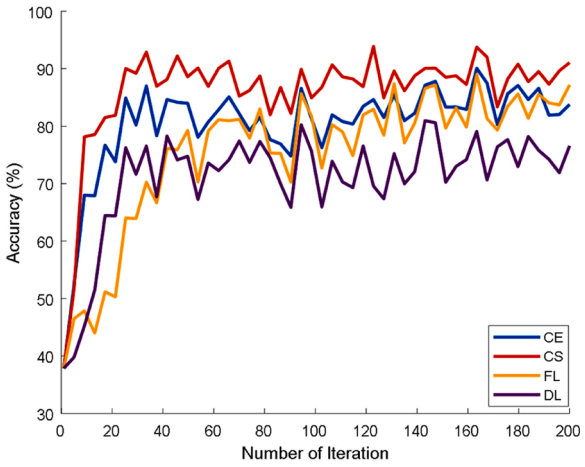


Fig. 6. The training accuracy of DiSegNet-2 using different loss functions including cross-entropy (CE) loss, cosine-sine (CS with $\gamma = 10$ and $\alpha = 0.64$) loss, focal loss (FL with $\gamma = 2$), and Dice loss (DL).

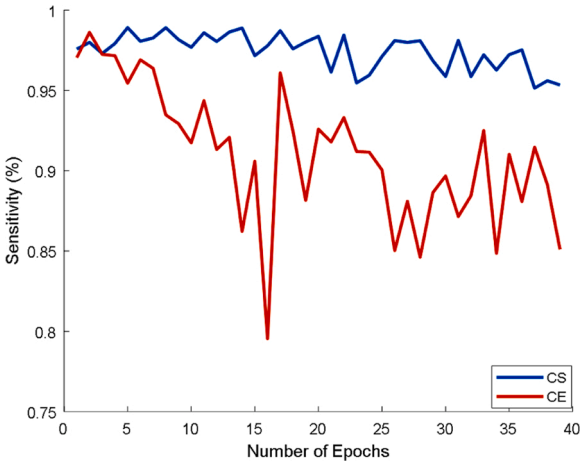


Fig. 7. The mean sensitivity after different epochs on one-fold 10 testing data sets.

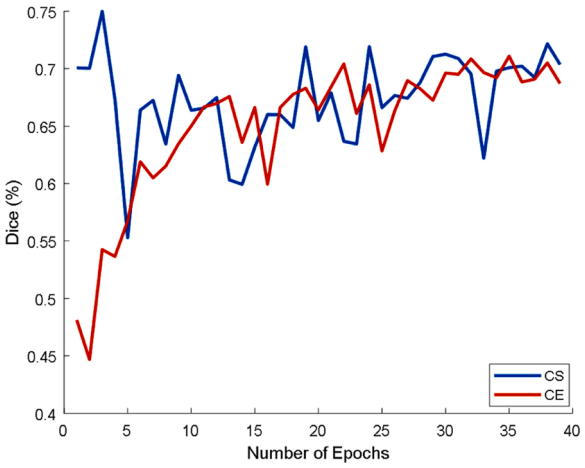


Fig. 8. The mean DSC after different epochs on one-fold testing data sets.

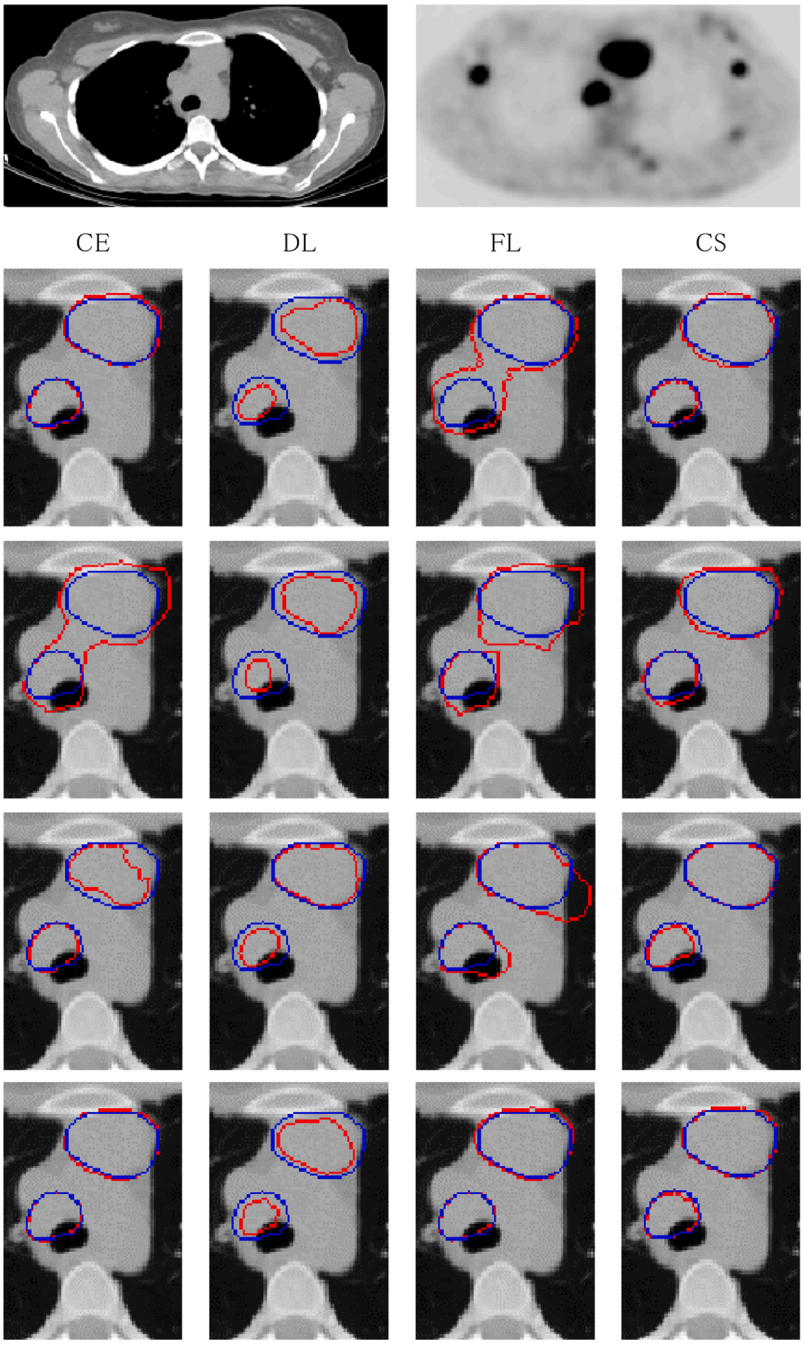


Fig. 9. shows one single representative axial slice of the LN segmentation results by using different networks with different loss functions.

architecture of DiSegNet-2 is seen in Fig. 5(b) where we put the two ASPP modules into the last convolution layers after concatenation to the feature maps from the last up-sampling layer. It should be noted that we use 1×1 convolution and transpose convolution after ASPP to make the size of feature map equal to the output mask.

Dice similarity coefficient (DSC) is utilized to evaluate the performance of the LN segmentation in terms of voxel level and region level. The DSC is calculated as follows:

$$DSC = \frac{2TP}{2TP + FP + FN} \tag{6}$$

The four-fold cross validation strategy is employed with 10 mutually exclusive data sets in each fold used for testing, namely leave-10-cases-out validation. Note that we did not use validation set for early termination of optimization. The mean and standard deviation (SD) of DSC under different architectures are shown in Table 1 below.

Comparing the different loss functions within the same network architecture (every row in Table 1), the proposed CS loss function achieved the best result in terms of DSC. The network architecture that used multi-stage Atrous-spatial-pyramid-pooling (MS-ASPP) achieves good performance of DSC compared to other network architectures, indicating that the MS-ASPP added to the SegNet architecture could reduce false positive voxels while maintaining a good performance for true positive voxels.

We used the first 200 iterations of the DiSegNet-2 to compare the training performance using different loss functions. The accuracy of the training stage was assessed to see the changes based on the various loss functions as seen in Fig. 6 below.

The accuracy was the highest with use of the CS loss function in the training stage, which means that the time for training for the neural

network can be reduced by using the CS loss function. Experiments were also performed by calculating the testing performance of one fold in different epochs, where an epoch refers to a unit of iterations throughout the whole training set. Here, we compare SEN (Sensitivity) and DSC (Dice similarity coefficient) for CE and CS as the loss functions by using the architecture of DiSegNet-2 as seen in Figs. 7 and 8below. SEN was much more stable by using the CS loss function compared to the CE loss function. Meanwhile, the performance as per DSC is better with CS than with CE.

Fig. 9: The first row shows single CT and PET image slices along with abnormally enlarged FDG-avid lymph nodes in the mediastinum (pre-vascular and right paratracheal stations) and bilateral axillae. The abnormal mediastinal lymph node segmentation results using SegNet, FCN16 s, DiSegNet-1, and DiSegNet-2 are shown in the second through fifth rows, respectively. The blue contours indicate the ground truth boundaries of selected mediastinal lymph nodes, and the red contours indicate segmentation results. The segmentation results utilizing CE, DL, FL, and CS loss functions are shown in the first column through fourth columns, respectively.

The segmentation results following use the proposed CS loss function were more stable comparing to CE, DL and FL loss function, which demonstrated in the last column. Moreover, the results from the DiSegNet which used MS-ASPP are more competitive, especially the DiSegNet-2. In summary, the DiSegNet with CS loss function can achieve better performance than other combinations, which can be seen in the Fig. 9.

In Table 2, we compare the results of our approach with other published methods for automatic LN segmentation. To the best of our knowledge, there are no previously published works regarding the automatic pathological LN segmentation on PET/CT studies using DCNNs in thorax. Moreover, it is worth noting that the CT data sets from PET/CT are typically acquired as low dose unenhanced CT images. In summary, our approach does segmentation of pathologic LNs of thorax on low dose CT and PET images. The previous methods in (Bouget et al., 2019) and (Tang et al., 2019; Barbu et al., 2012b; I. N. B et al., 2003; Moe et al., 2019) did not make use of multi-scale information to help LNs segmentation. Yet, our approach integrate Atrous-spatial-pyramid-pooling (ASPP) module into SegNet architecture with novel CS loss function, which shows that it can help LNs segmentation.

4. Discussion

In this paper, we aim two issues, which are not addressed by the previous LNs segmentation. The first is the imbalance of voxel classes and the second is the lack of multi-scale information from feature map of SegNet. Keeping this in mind, we proposed a novel deep convolutional neural network (DCNN) named DiSegNet to overcome these two issues.

Table 2
Comparison with other methods on CT or PET/CT images for automatic lymph node segmentation. The standard deviation values are shown in parentheses. DSC = Dice similarity coefficient.

Method	Body Region	DSC	Data sets
Bouget et al. (2019)	Thorax	0.409 ± 9.67	15 lung cancer CT data sets
Tang et al. (2019)	Thorax + Abdomen	0.825 ± 0.112	176 lymphadenopathy CT data sets
Barbu et al. (2012b)	Axilla	0.80 ± 0.126	131 lymphadenopathy CT data sets
	Pelvis + Abdomen	0.76 ± 0.127	54 lymphoma CT data sets
Nogues et al. (I. N. B et al., 2003)	Thorax + Abdomen	0.82 ± 0.096	171 lymphadenopathy CT data sets
(Moe et al. (2019)	Head and neck	0.75 ± 0.12	197 tumor and pathologic lymph nodes PET/CT data sets
Ours (DiSegNet-1 with CS)	Thorax	0.77 ± 0.05	63 lymphoma PET/CT data sets

Firstly, the proposed CS loss function could up-weight the loss from misclassified voxels. The idea is similar to the AdaBoost(Freund and Schapire, 1995), which emphasize the misclassified examples (voxels) and try to improve the performance in the next training iteration. It makes the neural network focus on the misclassified voxels. Meanwhile, the CS loss function will also down-weight the loss from well-classified voxels compared to CE loss function. It will reduce the relative loss for well-classified examples, giving more focus on misclassified examples further. Compared the performance by using FL, which only down-weight the well-classified examples, the proposed CS loss function enables training highly accurate and fast DCNN for LNs segmentation in the presence of large number of background examples (Non LN voxels). Considering the location prior of LNs, for example, LNs cannot be inside any organ in the mediastinum, some anatomical structures can be excluded, which could relieve the imbalance of training classes in the further way.

Second, we integrate the module atrous spatial pyramid pooling (ASPP) into SegNet architecture which used multiple parallel atrous convolutional layer with different sampling rates. The multi-scale feature information could be extracted from feature maps by using ASPP, which will increase feature resolution to help LNs segmentation in the decoder part of the network. Here, we designed two ways that integrate the ASPP into SegNet (named DiSegNet-1 and DiSegNet-2) in order to explore the optimal way to make use of multi-scale feature information from dilated operation. Compared to the final segmentation results, we found that the DiSegNet-1 can achieve better DSC than DiSegNet-2. The main reason may lie in the mode of the multi-scale information organization. In the DiSegNet-1, the multi-scale feature maps passed to the each layer of the decoder part, which may contain more information for improving the DSC performance.

Despite the novelties discussed above, there are also some limitations in this work. Firstly, we did not explore the performance that trained by single modalities, e.g., CT or PET. Second, the shape and location information are of LN have not integrated into the proposed DiSegNet architecture, which may help LNs segmentation task.

5. Conclusion

In this work, we proposed a simple but effective cosine-sine (CS) loss function as an objective function for training different networks to deal with the imbalance class problem for LN segmentation. The CS loss function can focus on learning the hard-to-classify (misclassified) voxels (examples) and down-weight the well-classified voxels (examples) at the same time. Our experimental results show that the proposed loss function can achieve good results for automatic abnormal LN segmentation in PET/CT images.

Moreover, we designed a multi-stage Atrous spatial pyramid pooling (MS-ASPP) sub-module that can be integrated into the SegNet architecture to improve the semantically accurate predictions and detailed segmentation along LN boundaries owing to the ability of multi-scale feature learning.

The proposed CS loss function and DiSegNet can also be used in other applications, such as natural images, and the encoder module can be replaced by using other network structures such as ResNet, or AlexNet, which can be studied in future research.

CRediT authorship contribution statement

Guoping Xu: Investigation, Software, Writing - original draft. **Hanqiang Cao:** Methodology, Validation. **Jayaram K. Udupa:** Supervision, Conceptualization, Methodology, Writing - review & editing. **Yubing Tong:** Visualization, Writing - review & editing. **Drew A. Torigian:** Data curation, Resources, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledges

The training of Mr. Guoping Xu in the Medical Image Processing Group, Department of Radiology, University of Pennsylvania, Philadelphia, for the duration of one year was supported by the China Scholarship Council. His subsequent training was supported by research funds provided by Dr. Drew Torigian at the University of Pennsylvania.

References

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39 (12), 2481–2495.

Barbu, A., Suehling, M., Xu, X., Liu, D., Zhou, S.K., Comaniciu, D., 2012a. Automatic detection and segmentation of lymph nodes from CT data. IEEE Trans. Med. Imaging 31 (2), 240–250.

Barbu, A., Suehling, M., Xu, X., Liu, D., Zhou, S.K., Comaniciu, D., 2012b. Automatic detection and segmentation of lymph nodes from CT data. IEEE Trans. Med. Imaging 31 (2), 240–250.

Bouget, D., Jørgensen, A., Kiss, G., Leira, H.O., Langø, T., 2019. Semantic segmentation and detection of mediastinal lymph nodes and anatomical structures in CT data for lung cancer staging. Int. J. Comput. Assist. Radiol. Surg. 14 (6), 977–986.

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848.

Eigen, D., Fergus, R., 2015. In: Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture. Proceedings of the IEEE international conference on computer vision, pp. 2650–2658.

Feulner, J., Kevin Zhou, S., Hammon, M., Hornegger, J., Comaniciu, D., 2013. Lymph node detection and segmentation in chest CT data using discriminative learning and a spatial prior. Med. Image Anal. 17 (2), 254–270.

Freund, Y., Schapire, R.E., 1995. A desicion-theoretic generalization of on-line learning and an application to boosting. European Conference on Computational Learning Theory 23–37.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 770–778.

Hoogi, A., Lambert, J.W., Zheng, Y., Comaniciu, D., Rubin, D.L., 2017. A Fully-Automated Pipeline for Detection and Segmentation of Liver Lesions and Pathological Lymph Nodes. arXiv preprint arXiv:1703.06418.

I. N. B, et al., 2003. In: International Conference on Medical Image Computing and Computer-Assisted InterventionAutomatic lymph node cluster segmentation using holistically-nested neural networks and structured optimization in CT images, 2878, pp. 388–397.

Kaiming, H., Xianyu, Z., Ren, S., Sun, J., 2015. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. Proceedings of the IEEE International Conference on Computer Vision 1026–1034.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, pp. 1097–1105.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollar, P., 2017. Focal loss for dense object detection. Proceedings of the IEEE International Conference on Computer Vision 2999–3007.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully Convolutional Networks for Semantic Segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Moe, Y.M., et al., 2019. Deep learning for automatic tumour segmentation in PET/CT images of patients with head and neck cancers. arXiv Prepr. arXiv1908.00841.

Oda, H., et al., 2018. Dense volumetric detection and segmentation of mediastinal lymph nodes in chest CT images. Medical Imaging 2018: Computer-Aided Diagnosis. International Society for Optics and Photonics 10575, 1057502.

Ren, S., He, K., Girshick, R., Sun, J., 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. (6), 1137–1149.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. Int. Conf. Med. image Comput. Comput. Interv. 234–241.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556.

Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J., 2017. Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. Deep learning in medical image analysis and multimodal learning for clinical decision support, pp. 240–248.

Szegedy, C., et al., 2015. Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 1–9.

Tang, Y., Oh, S., Xiao, J., Summers, R.M., Tang, Y., 2019. CT-realistic data augmentation using generative adversarial network for robust lymph node segmentation. Medical Imaging 2019: Computer-Aided Diagnosis. International Society for Optics and Photonics, 10950: 109503V.

Udupa, J.K., et al., 2014. Body-wide hierarchical fuzzy modeling, recognition, and delineation of anatomy in medical images. Med. Image Anal. 18 (5), 752–771.

Xu, G., et al., 2018. Thoracic lymph node station recognition on CT images based on automatic anatomy recognition with an optimal parent strategy, Medical Imaging 2018: Image Processing. Int. Soc. Optics and Photonics 10574, 105742F.

Xu, G., et al., 2020. A novel exponential loss function for pathological lymph node image classification. MIPPR 2019: Parallel Process. Images and Optimization Techniques; and Medical Imaging 11431, 114310A.