Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

DSU-Net: Distraction-Sensitive U-Net for 3D lung tumor segmentation

Junting Zhao [a], Meng Dang [a], Zhihao Chen [a], Liang Wan [a,b,*]

[a] *College of Intelligence and Computing, Tianjin University, Tianjin, China*
[b] *Medical College, Tianjin University, Tianjin, China*

ARTICLE INFO

Keywords:
Lung tumor
3D-segmentation
CT volumes
U-Net
Distraction attention

ABSTRACT

Automatic segmentation of lung tumors is a crucial and challenging problem. Many existing methods suffer from ambiguity of tissue regions and tumor regions, which occur with similar appearance. To address this problem, we propose a new cascaded two-stage U-net model, Distraction-Sensitive U-Net (DSU-Net), to explicitly take the ambiguous region information (referred as distraction region) into account. Stage-I generates a global segmentation for the whole input CT volume and predicts latent distraction regions, which contain both false negative areas and false positive areas, against the segmentation ground truth. Stage-II embeds the distraction region information into local segmentation for volume patches to further discriminate the tumor regions. To this end, a Distraction Attention Module (DAM) is proposed and applied in each level of U-Net in Stage-II, to improve the discrimination of features. We evaluate our network on a lung cancer dataset from Gross Target Volume segmentation of MICCAI2019 challenge. Experimental results show that the proposed DSU-Net outperforms existing U-like networks.

1. Introduction

As one of the most common human cancers, lung cancer is one leading cause of death among cancer-related diseases (Liu et al., 2018). About 2 million new cases of lung cancer occur every year over the world, with a mortality rate of 20% (Suster and Mino-Kenudson, 2020). Having a way for quickly locating malignant lung tumors and accurately delineating them is critical for many clinical applications, including radiological diagnosis, treatment planning, etc. Currently, several imaging techniques are used for lung cancer, i.e. computed tomography (CT), positron emission tomography (PET), ultrasound, etc, among which CT is the most widely used means. The traditional way to analyze lung cancer CT images demand radiologists and/or physicians to manually mark the locations, sizes, and other information of lung tumors. However, the large amount of volume slices in CT images makes the interpretation and analysis process time-consuming, and also easily leads to human bias and mistakes.

Computer-aided techniques for automatic medical image segmentation have been developed rapidly, which can help to speed up the manipulation of medical images, lighten the workload of radiologists and physicians, and also assist doctors to improve the accuracy of disease diagnosis (Shen et al., 2017). Early automatic lung tumor segmentation techniques include threshold based methods (Taheri et al., 2010; Sujji et al., 2013), active contour based methods (Yazdanpanah et al., 2009), region based methods (Pan and Lu, 2007; Sato et al., 2002), conditional random field based methods (Hu et al., 2008), etc. Later on, data driven learning based techniques, e.g. support vector machine based methods (Netto et al., 2012; Asuntha et al., 2016), random forest based methods (Zhao et al., 2017), become the mainstream in the literature, showing better performance with higher accuracy and efficiency.

In recent years, with the development of deep learning techniques, neural network based methods raise a lot of attentions. For instance, Pratiksha et al. used deep learning based method to find the range of the lung texture pattern of diseases from CT images (Pratiksha, 2017). Chen et al. proposed a hybrid segmentation network (HSN) which combines a lightweight 3D CNN to learn long-range 3D contextual information and a 2D CNN to learn fine-grained semantic information (Chen et al., 2019). Li et al. segmented lung tumor using a variational method on PET-CT images (Li et al., 2020). They first applied a 3D Fully Convolutional Network (FCN) on the CT image to produce a probability map, then proposed a fuzzy variational model to incorporate the probability map and the PET intensity image for an accurate segmentation.

Notice that many deep learning based segmentation methods build upon U-Net (Ronneberger et al., 2015), which is the mostly used encoder–decoder architecture for medical image segmentation. Thanks to its skip-connection structure, U-Net can use the prediction information at low layers to improve the quality of segmentation prediction results. Till now, many U-like networks have been proposed for medical image segmentation, such as 3D U-Net (Çiçek et al., 2016), nnU-Net (Isensee et al., 2018), Attention U-Net (Oktay et al., 2018),

[*] Corresponding author at: College of Intelligence and Computing, Tianjin University, Tianjin, China.
E-mail addresses: zhaojt@tju.edu.cn (J. Zhao), dangmeng@tju.edu.cn (M. Dang), zh_chen@tju.edu.cn (Z. Chen), lwan@tju.edu.cn (L. Wan).

https://doi.org/10.1016/j.engappai.2021.104649
Received 19 August 2021; Received in revised form 15 November 2021; Accepted 26 December 2021
Available online 11 January 2022

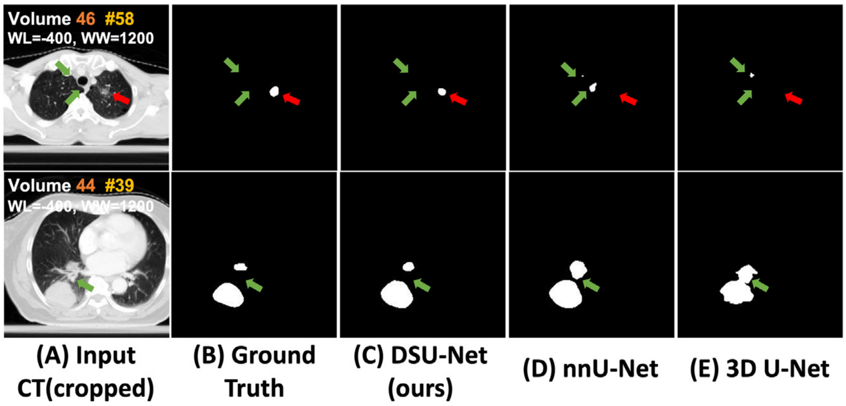

Fig. 1. Lung tumor segmentation results from existing methods (nnU-Net and 3D U-Net) and our method. Volume 46 and volume 44 are two patients' CT volumes. WL and WW are short for window level and window width, respectively. $\#X$ denotes the Xth slice along the transverse axis. Green arrows point to non-tumor regions that appear like tumors, and red arrows point to tumor regions that appear like non-tumors. We can see that nnU-Net and 3D U-Net wrongly detect these regions.

R2U-Net (Alom et al., 2018), UNet++ (Zhou et al., 2018), etc., which report good segmentation results in various medical scenarios. In the application of lung tumor segmentation, we experiment with 3D U-Net and nnU-Net, which are two well-known U-like models. As shown in Fig. 1(D)(E), nnU-Net and 3D U-Net may wrongly detect the non-tumor regions in the image that appear like tumors (pointed by green arrows), which we refer as false negative distraction regions, and also wrongly detect the tumor regions that appear like non-tumor regions (pointed by red arrows), which is referred as false positive distraction regions. The reason for this problem is that the distraction regions are ambiguous regions and they are handled equally as other regions.

To address this problem, we propose a new Distraction-Sensitive U-Net (DSU-Net) to explicitly take the ambiguous region information into account. The DSU-Net follows a two-stage coarse-to-fine structure. Stage-I takes the whole resized image as input and outputs the global segmentation result. We then compare the global result with the ground truth to obtain distraction regions. The output of stage-I is concatenated with the original image and then cropped as patches to be fed into stage-II. Stage-II embeds the distraction region information in each level of U-Net to refine the segmentation. In this stage, we propose a Distraction Attention Module (DAM) to enhance the false negative regions and suppress the false positive regions, by considering extra supervisions obtained at stage-I. The contributions of our work can be summarized as follows:

(1) We consider the impact of distraction regions on tumor segmentation, and propose a new two-stage cascaded U-like model, DSU-Net, for 3D lung tumor segmentation.
(2) We design a distraction attention module (DAM) to improve the discrimination of features, especially on those distraction regions. According to our knowledge, this is the first work to introduce distraction semantics in medical image segmentation tasks.
(3) On Gross Target Volume segmentation of lung cancer dataset, the experiments demonstrate the superiority of our method over those U-like methods for lung cancer segmentation, especially for ambiguous regions.

2. Related work

2.1. Traditional methods for CT medical image segmentation

Manual and semi-manual techniques for medical image segmentation of CT images are usually subjective, error-prone, operator-dependent, and easily lead to a heavy workload for physicians and radiologists. The early automatic medical image segmentation methods contain morphological operations based method (Kostis et al., 2003), conditional random field based method (Hu et al., 2008), region growth based method (Dehmeshki et al., 2008), active contour based method (Chan and Vese, 2001; Reboucas et al., 2011; Reboucas Filho et al., 2017) and etc. However, these hand-craft feature-based methods are not widely applied in clinics, because they are usually slow and not robust enough to heterogeneous, low-contrast CT images in real-life (Wang et al., 2020).

2.2. Deep learning based methods for lung tumor segmentation

In recent years, deep learning methods have been widely used for medical image segmentation with highly competitive results compared to traditional methods. Xu et al. introduced a deep convolutional neural network to segment lungs with complex opacities on CT images (Xu et al., 2017). Pratiksha Hattikatti proposed a CNN for finding the range of the lung texture pattern of diseases from CT images (Pratiksha, 2017). Zhong et al. used FCN for 3D tumor segmentation based on PET-CT dual-modality images (Zhong et al., 2018). Later on, considering that the medical image dataset is usually small, Li et al. proposed to first produce a probability map from the CT image via a 3D FCN, and then applied fuzzy variational model incorporating the probability map and the PET intensity image for accurate tumor segmentation (Li et al., 2020). Chen et al. proposed a hybrid segmentation network, which combines a lightweight 3D CNN to learn long-range 3D contextual information and a 2D CNN to learn fine-grained semantic information (Chen et al., 2019).

2.3. Two-stage based methods on medical image segmentation

In this section, we focus on multiple applications of two-stage based methods for medical images. Chang and Teng presented a two-stage self-organizing map approach, which can precisely identify dominant color components to discover the region of interest for diagnosis purposes (Chang and Teng, 2007). Roth et al. claimed a two-stage, coarse-to-fine approach targeting three anatomical organ segmentation (liver, spleen, pancreas). It first uses a 3D FCN to roughly define a candidate region and then uses the region as input to the second 3D FCN, which focuses on more detailed segmentation of the organs and vessels (Roth et al., 2018). A fully automated two-stage framework for pancreas segmentation was proposed by Zhao et al. (2019). The first stage trained a U-Net for the down-sampled 3D volume segmentation to produce candidate regions, and then another 3D U-Net is trained on the candidate regions which possibly cover the pancreas. In the study of Cao et al. (2019), they claimed a two-stage CNN for lung nodule detection. In the first stage, a U-Net with a new sampling training strategy achieves an initial detection of lung nodules. The second stage utilizes a dual pooling structure to reduce false positives. Liu et al. used a two-stage U-Net framework and an adaptive threshold window to automatically segment the whole heart and heart substructures (Liu et al., 2019). The first stage extracts the regions of interest from the whole heart, while the second stage is to segment the various substructures. An adaptive threshold window is used to remove the noisy parts of the data while preserving the anatomical relationships between local regions.

Note that most of the above methods perform the overall segmentation in the first stage and then the second stage conducts the partial segmentation. In our approach, stage-I and stage-II both segment the target regions, in a coarse to fine manner. Moreover, the output of stage-I is closely coupled with stage II, in which it not only used as the part of the input of stage-II, but also used as a supervision on stage-II.

2.4. Distraction attention

Distraction concepts have been explored in many computer vision tasks, such as semantic segmentation (Huang et al., 2017), saliency detection (Chen et al., 2020; Xiao et al., 2018) and visual tracking (Zhu et al., 2018). Most of existing methods suppress negative high-level representations or filter out the distracting regions (Xiao et al., 2018; Huang et al., 2017; Chen et al., 2020; Zhu et al., 2018). Instead, Zheng et al. used distraction cues to improve the performance of shadow distraction (Zheng et al., 2019). They split shadow distraction into false negative estimates and false positive estimates, and designed specific architectures to efficiently integrate the two types of distraction semantics. Inspired by this method, we developed a new Distraction Attention Module to explicitly learn semantic features of the distraction regions, and embedded these features in U-Net structure. To the best of our knowledge, it is the first work to introduce distraction semantics in medical image segmentation tasks.

3. DSU-net model

3.1. Network architecture

As shown in Fig. 2, we propose a Distraction-Sensitive U-Net model, shortly named DSU-Net, by integrating nnU-Net (Isensee et al., 2018) and our designed Distraction-Attention Module (DAM). Our DSU-Net inherits nnU-Net's coarse-to-fine structure, which consists of two cascaded U-Nets. The first U-Net processes the resized CT images to cover the rich context information (stage-I). The second U-Net processes the cropped images to obtain the high resolution segmentation prediction (stage-II). To build the connection between global and local aspects, the predictions of the first U-Net is fed into the second U-Net via concatenating with the original input images. This two-stage coarse-to-fine strategy is effective for many 3D pixel-level tasks, especially when dealing with small objects. However, it still suffers from the ambiguity of tumor regions and tissue regions with similar appearance. To further enhance the discriminating ability for the distraction regions, we propose to embed distraction information, detected from stage-I, into each level of stage-II.

Specifically, we denote the input image by $\mathbf{I} \in \mathbb{R}^{1\times D\times H\times W}$, where D, H and W represent the depth, height and width of CT volume, respectively. Taking the nnU-net as the backbone, the corresponding prediction of nnU-Net, $\mathbf{P}_n \in \mathbb{R}^{1\times D\times H\times W}$, can be formulated as follow:

$$\mathbf{P}_n = \mathcal{F}_r(\mathbf{G}_{im} \mid \mathtt{Cat}(\mathbf{I}, \mathcal{F}_c(\mathbf{G}_{im} \mid \mathbf{I}, \theta_c)), \theta_r), \tag{1}$$

where $\mathbf{G}_{im} \in \mathbb{R}^{1\times D\times H\times W}$ denotes the ground-truth of the input volume, $\mathcal{F}_c(\mathbf{G}_{im}|\cdot, \theta_c)$ represents the coarse U-Net with parameters θ_c under the supervision of $\mathbf{G}_{im}$, $\mathcal{F}_r(\mathbf{G}_{im}|\cdot, \theta_r)$ represents the fine U-Net, and $\mathtt{Cat}(\cdot)$ is the concatenation operator.

Inspired by shadow distraction work (Zheng et al., 2019), we think that if the network is encouraged to simultaneously predict the tumor regions and the possible distraction regions, the network will have stronger discriminating power in these regions, and hence can improve the segmentation performance for the whole volume. Following this idea, we need to determine which regions are distraction regions and explore how to use them to guide segmentation. A direct way of getting distraction regions is to collect the average prediction of different networks (Zheng et al., 2019). However, this strategy uses extra training of many networks. Thanks to the two-stage architecture of the nnU-Net, in our work, we can use the results of stage-I to estimate the distraction regions.

Given the input image $\mathbf{I}$ and the ground-truth $\mathbf{G}_{im}$, the false positive supervision $\mathbf{G}_{fp} \in \mathbb{R}^{1\times D\times H\times W}$ and the false negative supervision $\mathbf{G}_{fn} \in \mathbb{R}^{1\times D\times H\times W}$ can be formulated as:

$$\mathbf{G}_{fn} = \mathbf{G}_{im} \ / \ (\mathcal{F}_c(\mathbf{G}_{im} \mid \mathbf{I}, \theta_c)), \tag{2}$$

$$\mathbf{G}_{fp} = \mathcal{F}_c(\mathbf{G}_{im} \mid \mathbf{I}, \theta_c) \ / \ \mathbf{G}_{im}, \tag{3}$$

where the operator / is used to obtain difference set between two images, e.g., $\mathbf{A}$ / $\mathbf{B}$ represents the set of elements in $\mathbf{A}$ but not in $\mathbf{B}$.

We then embed the distraction information in each level of decoder in stage-II via the proposed DAMs, to make our model sensitive to distraction regions. This deep supervision is proven to be an effective way for more stable training (e.g., Long et al., 2015; Chen et al., 2017; Badrinarayanan et al., 2017). The final prediction $\mathbf{P}_r$ of our DSU-Net can be obtained by:

$$\mathbf{P}_r = \mathcal{F}_r(\mathbf{G}_{im}, \mathbf{G}_{fn}, \mathbf{G}_{fp} \mid \mathtt{Cat}(\mathbf{I}, \mathcal{F}_c(\mathbf{G}_{im} \mid \mathbf{I}, \theta_c)), \theta_r), \tag{4}$$

where $\mathcal{F}_r(\mathbf{G}_{im}, \mathbf{G}_{fn}, \mathbf{G}_{fp}|\cdot, \theta_r)$ represents the fine U-Net (i.e., stage-II) with parameters θ_r under the supervisions of $\mathbf{G}_{fn}, \mathbf{G}_{fp}$ and $\mathbf{G}_{im}$.

Compared with the original nnU-Net, our DSU-Net focuses more on the distraction regions in stage-II. By collecting the results of stage-I and using it to guide the training of stage-II, we build a new bridge between the two stages other than the original sequential connection. Since stage-I and stage-II consider the global and local aspects of the input volume, respectively, our DSU-Net benefit from these two aspects naturally.

3.2. Distraction attention module

In order to increase the discriminating power in the distraction regions, we design the Distraction-Attention Module (DAM) in stage-II, to allow the network to learn the characteristics of G_{fn} and G_{fp} explicitly. DAM contains two attention submodules, as illustrated in Fig. 3(A). (1) The first is false negative attention module, shortly named FNAM, which enhances the false negative aspects in input $\mathbf{F}$. (2) The second is false positive attention module, shortly named FPAM, which suppresses the occurrence of false positive regions.

We denote the input feature in each level (scale) by $\mathbf{F} \in \mathbb{R}^{C_i\times D_i\times H_i\times W_i}$, where (C_i, D_i, H_i, W_i) is feature size at the ith scale. The output feature is denoted as $\hat{\mathbf{F}} \in \mathbb{R}^{C_i\times D_i\times H_i\times W_i}$, which has the same size as the input $\mathbf{F}$. Passing through these two sub-modules, we can get the output feature $\hat{\mathbf{F}}$ as follow:

$$\hat{\mathbf{F}} = \mathbf{F} + \mathbf{F}_{fne} - \mathbf{F}_{fpe}, \tag{5}$$

where

$$\mathbf{F}_{fne} = \mathcal{F}_{FNAM}(\mathbf{F}, \theta_n), \tag{6}$$

$$\mathbf{F}_{fpe} = \mathcal{F}_{FPAM}(\mathbf{F}, \theta_p). \tag{7}$$

Here, $\mathcal{F}_{FNAM}(\cdot, \theta_n)$ and $\mathcal{F}_{FPAM}(\cdot, \theta_p)$ represent the FNAM with parameters θ_n and FPAM with parameters θ_p, respectively. $\mathbf{F}_{fne} \in \mathbb{R}^{C_i\times D_i\times H_i\times W_i}$ and $\mathbf{F}_{fne} \in \mathbb{R}^{C_i\times D_i\times H_i\times W_i}$ are intermediate products of these two sub-modules. It is worth noting that we use the residual-like structure (i.e., single element-wise plus and minus operations) to enhance or suppress the input features. As pointed in He et al. (2016), this residual-like structure can accelerate convergence and ensure that the $\hat{\mathbf{F}}$ is not worse than $\mathbf{F}$.

As for the sub-modules FNAM and FPAM, we design the same architectures as shown in Fig. 3(B), while they have different supervisions.

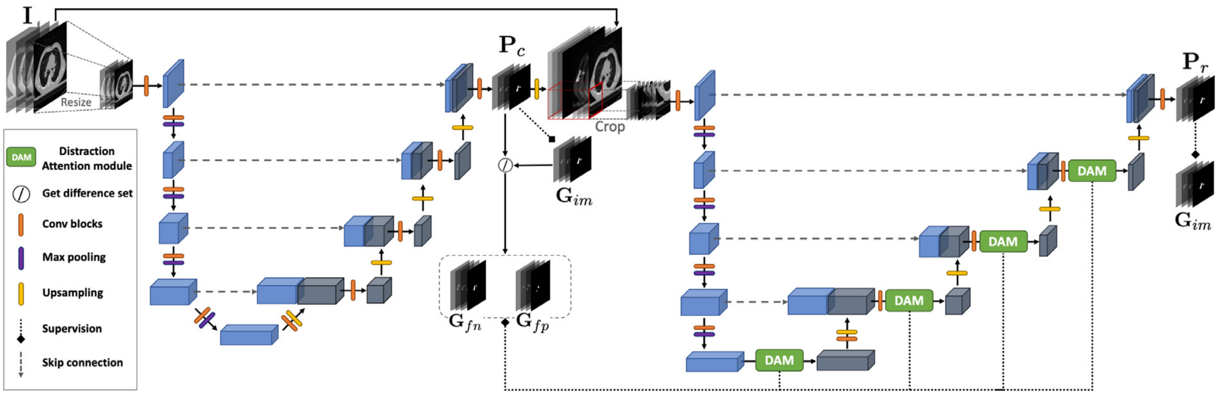

Fig. 2. The architecture of our proposed Distraction-Sensitive U-Net (DSU-Net). The whole image (**I**) is resized and fed into the stage-I to produce the prediction ($\mathbf{P}_c$) which covers the rich context information. Then $\mathbf{P}_c$ is concatenated with the original image (**I**) and together cropped to obtain the high resolution segmentation prediction ($\mathbf{P}_r$) in stage-II. The false positive supervision ($\mathbf{G}_{fp}$) and the false negative supervision ($\mathbf{G}_{fn}$) are calculated by comparing the ground-truth ($\mathbf{G}_{im}$) and the stage-I result ($\mathbf{P}_c$). DAMs are added on each level of the decoder in stage-II, to guide each level with distraction-sensitive information.

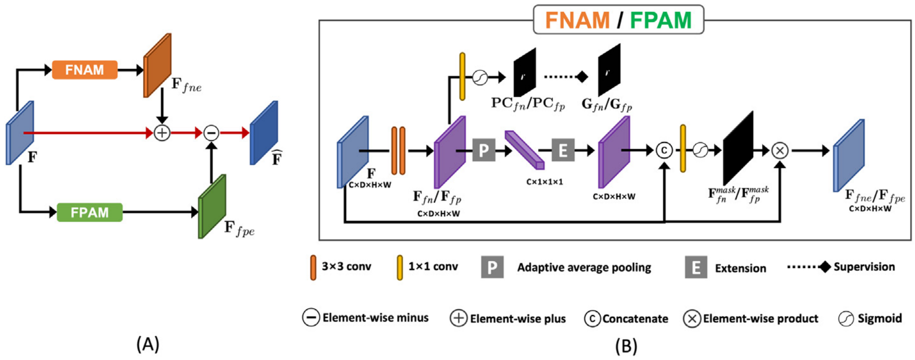

Fig. 3. The architecture of our Distraction Attention Module (DAM) in DSU-Net. **(A)** DAM contains two submodules, FNAM and FPAM. **(B)** FNAM and FPAM have the same architectures, yet with different supervision from $\mathbf{G}_{fn}$ and $\mathbf{G}_{fp}$.

For convenience, we take the FNAM as an example. FNAM is designed to learn false negative features $\mathbf{F}_{fn} \in \mathbb{R}^{C_i \times D_i \times H_i \times W_i}$ and produces false negative enhanced features $\mathbf{F}_{fne}$. We first employ a feature extractor composed of two 3×3 convolutions to extract the $\mathbf{F}_{fn}$. To force the $\mathbf{F}_{fn}$ to capture the underlying semantics, we add supervision to $\mathbf{F}_{fn}$. Here, we take the difference between the output of stage-I and the ground truth as the supervision. On the other hand, to better gain the global information of false negative regions, we utilize the adaptive average pooling and extension operations to scale features. After these two operations, the new features $\mathbf{F}'_{fn}$ are concatenated with the input feature $\mathbf{F}$, and fed into an attention block to produce a soft mask, $\mathbf{F}^{mask}_{fn} \in [0,1]^{1 \times D_i \times H_i \times W_i}$. The masked feature $\mathbf{F}_{fne}$ is obtained by multiplying $\mathbf{F}$ with duplicated $\mathbf{F}^{mask}_{fn}$ (along feature channel) element-wisely. Finally, we get the $\mathbf{F}_{fne}$ as follow:

$$\mathbf{F}_{fne} = \mathbf{F} \odot \sigma(\mathtt{Conv}(\mathtt{Cat}(\mathbf{F}'_{fn}, \mathbf{F}))), \tag{8}$$

where $\odot$ denotes element-wise product, σ is the `Sigmoid` operation, and `Conv` is 1×1 convolutions for reshape features. Similarly, the false positive enhanced features can be computed as:

$$\mathbf{F}_{fpe} = \mathbf{F} \odot \sigma(\mathtt{Conv}(\mathtt{Cat}(\mathbf{F}'_{fp}, \mathbf{F}))). \tag{9}$$

3.3. *Loss functions*

To better handle the scale variance of lung tumors, we fuse the binary cross entropy (BCE) loss (Eq. (12)) with the Dice coefficient loss (Eq. (11)) to compute the tumor segmentation loss $\boldsymbol{\Phi}$,

$$\Phi(\mathbf{Y}^*, \mathbf{Y}) = \Phi_{BCE} + \Phi_{Dice}, \tag{10}$$

where,

$$\Phi_{Dice} = 1 - \frac{2TP}{2TP + FP + FN}, \tag{11}$$

$$\Phi_{BCE} = -\frac{1}{N} \sum_{p \in N} \left(\mathbf{Y}^*_p \log \mathbf{Y}_p + (1 - \mathbf{Y}^*_p) \log(1 - \mathbf{Y}_p) \right). \tag{12}$$

Here, $\mathbf{Y}^*$ and $\mathbf{Y}$ denote the target image and prediction image, respectively; N is the set of all voxels in the volume; $\mathbf{Y}^*_p$ and $\mathbf{Y}_p$ denote the values at voxel p in the prediction and the target images, respectively. TP, FP and FN denote truth positive, false positive and false negative, respectively.

The training process of our network is split into two parts. At stage-I, we train the coarse U-Net by optimizing the prediction of lung tumor, which minimizes the objective function:

$$\mathcal{L}^c = \Phi(\mathbf{G}_{im}, \mathbf{P}_c), \tag{13}$$

where $\mathcal{L}^c$ denotes the loss of the coarse U-Net, and $\mathbf{P}_c$ is the prediction of the coarse U-Net. At stage-II, we train the fine U-Net by optimizing the prediction of lung tumor, false positive and false negative at multi-scales, which minimizes the following objective function:

$$\mathcal{L}^r = \Phi(\mathbf{G}_{im}, \mathbf{P}_r) + \alpha \sum_{i=1}^{4} \Phi(\mathbf{G}_{fn}, \mathbf{P}^i_{fn}) + \beta \sum_{i=1}^{4} \Phi(\mathbf{G}_{fp}, \mathbf{P}^i_{fp}), \tag{14}$$

where $\mathcal{L}^r$ denotes the loss of the fine U-Net, and $\mathbf{P}_r$ is the prediction of the fine U-Net; $\mathbf{P}^i_{fn}$ and $\mathbf{P}^i_{fp}$ are predictions of the false negative and false positive regions at the ith scale, respectively. Here, α and β are hyper parameters to balance the training weight, for which we empirically set $\alpha = 1$, $\beta = 1$.

3.4. Optimization and inference

Optimization process. In the literature, there are many works discussing optimization methods, including expert system modeling (Pozna and Precup, 2014), artificial neural networks optimization (Nayak et al., 2018; Albu et al., 2019; Mishra et al., 2020), relational classifiers (Zall and Kangavari, 2019; Borlea et al., 2021), etc. To find optimal weights for our DSU-Net, we utilize SGD optimizer and optimize Stage-I and Stage-II in a separate manner.

Specifically, when we train Stage-I, the parameters of the coarse net $\mathbf{F_c}$, θ_c, are randomly initialized. Following the normal way used in existing related works, we apply kaiming-normal initialization for all convolutional layers, and use constant-normal initialization for all batch normalization layers. Then we optimize the θ_c by minimizing $\mathcal{L}^c$ in Eq. (13). Here, we set the start learning rate as 3e-4 and employ the learning rate decay strategy of ReduceLROnPlateau, which reduces the learning rate by a factor of 0.1 once the learning loss does not decrease during recent 50 epochs. We restrict the maximal epoch number as 1000. After optimizing Stage-I, we fix the θ_c, and start the training of Stage-II, i.e., $\mathbf{F_r}$. In this stage, we randomly initialize θ_r, similar to that of Stage-I, and optimize the θ_r by minimizing $\mathcal{L}^r$ in Eq. (14). The details of the optimization process are given in Algorithm 1.

Algorithm 1: Training process of DSU-Net.

```
Stage-I:
Input: Original input image $\mathbf{I}$; Ground Truth $\mathbf{G}_{im}$
Initialize $\theta_c$(parameters of coarse net $\mathbf{F}_c$) randomly;
repeat
    Compute Stage-I prediction: $\mathbf{P}_c = \mathbf{F}_c(\mathbf{I}, \theta_c)$;
    Compute Stage-I loss $\mathcal{L}^c$ by Eq. (13);
    Optimize $\theta_c$ by minimizing $\mathcal{L}^c$;
until converge;
Fixed $\theta_c$ then end the training of Stage-I

Stage-II:
Input: Original input image $\mathbf{I}$; Ground Truth $\mathbf{G}_{im}$
Inherit $\theta_c$(parameters of coarse net $\mathbf{F}_c$) based on the upper process;
Initialize $\theta_r$(parameters of fine net $\mathbf{F}_r$) randomly;
repeat
    Compute the false negative supervison $\mathbf{G}_{fn}$ by Eq. (2);
    Compute the false positive supervison $\mathbf{G}_{fp}$ by Eq. (3);
    Compute Stage-II final prediction $\mathbf{P}_r$ by Eq. (4);
    Compute Stage-II loss $\mathcal{L}^r$ by Eq. (14);
    Optimize $\theta_r$ by minimizing $\mathcal{L}^r$;
until converge;
Fixed $\theta_r$ then end the training of Stage-II
```

Algorithm 2: Testing process of DSU-Net.

```
Stage-I:
Input: Original input image $\mathbf{I}$
Inherit $\theta_c$(parameters of coarse net $\mathbf{F}_c$) based on the training
  process;
Compute Stage-I prediction:

  $\mathbf{P}_c = \mathbf{F}_c(\mathbf{I}, \theta_c)$;                (15)

Output: Stage-I's prediction $\mathbf{P}_c$

Stage-II:
Input: Original input image $\mathbf{I}$; $\mathbf{P}_c$ from Stage-I
Inherit $\theta_r$(parameters of fine net $\mathbf{F}_r$) based on the training process;
Compute Stage-II prediction:

  $\mathbf{P}_r = \mathbf{F}_r(\mathtt{Cat}(\mathbf{I}, \mathbf{P}_c), \theta_r)$;      (16)

Output: Stage-II's prediction $\mathbf{P}_r$
```

Inference process. The testing process is summarized in Algorithm 2, in which Stage-I and Stage-II are applied in a sequential order. In more detail, we use the network weights θ_c and θ_r learnt from the training process. A given testing image is fed into $\mathbf{F}_c$ to compute Stage-I's prediction $\mathbf{P}_c$ (Eq. (15)). Then the testing image is concatenated with $\mathbf{P}_c$, and fed into $\mathbf{F}_r$ to compute the final prediction $\mathbf{P}_r$ (Eq. (16)). Note that in the inference process, the distraction information are discarded as their related information have been embedded in the learning process.

3.5. Implementation details

Our method is implemented with PyTorch, following the protocol present in Isensee et al. (2018). In stage-I, the input volume is downsampled to (128 × 128 × 64), and in stage-II, the input volumes is cropped into patches with a size of (128 × 128 × 64), while the batch size is set as 2. All the experiments are conducted on the same platform: Intel Intel(R) Core(TM)i7-7800X @3.50 GHz and GTX 2080Ti. We will release the code upon the paper acceptance, with the problem link on github: https://github.com/CindyZJT/DSU-Net.git.

4. Experiment

In this section, we first introduce evaluation dataset and evaluation metrics (Section 4.1), then compare our results both quantitatively and qualitatively with existing U-like Nets methods (Section 4.2). Finally, we conduct ablation study to analyze the impact of main modules in the proposed method (Section 4.3).

4.1. Dataset and evaluation metrics

In this paper, we use the public dataset of Gross Target Volume segmentation of lung cancer in the MICCAI 2019 Challenge with link https://structseg2019.grand-challenge.org/Home/. We denote it as "LC Dataset". The LC Dataset contains GTV annotations of 50 lung cancer patients' CT scans. When we trained our architecture, the challenge did not publish the test dataset. Hence, we divide LC Dataset randomly into LC training dataset and LC testing dataset, with a ratio of 4:1. In other words, the training dataset contains 40 CT scans, and the testing dataset contains 10 CT scans.

To evaluate the performance of the proposed model, we follow the Challenge requirements to use the two widely-used segmentation evaluation metrics, i.e. Dice Similarity Coefficient (DSC) and 95% Hausdorff Distance. **The Dice metric** measures volumetric overlap between segmentation results and groundtruth annotations (Eq. (17)). **Hausdorff distance** measures the spatial distance between resulted segments and groundtruth. 95% HD can eliminate the impact of small outliers (Eq. (18)), and it is more sensitive to the divided boundary. We also adopted Jaccard, Recall and Precision as evaluation metrics. **The Jaccard index** measures the intersection over union between the segmented image and the groundtruth (Eq. (19)). TP represents true positive cases, TN means true negative cases, FP means false positive cases, and FN means false negative cases. **Recall rate** represents the ratio of the number of correctly classified positive cases to the actual number of true ground truth cases (Eq. (20)). **Precision** indicates the ratio of the number of correctly classified positive cases to the number of predicting positive cases (Eq. (21)). Here, $\mathbf{Y}^*$ and $\mathbf{Y}$ denote the groundtruth annotation and the prediction result.

$$Dice(\mathbf{Y}^*, \mathbf{Y}) \quad = \frac{2TP}{FP + 2TP + FN}, \tag{17}$$

$$\begin{aligned} d_H(\mathbf{Y}, \mathbf{Y}^*) &= max\left\{d_{\mathbf{YY}^*}, d_{\mathbf{Y}^*\mathbf{Y}}\right\} \\ &= max\left\{\max_{y \in \mathbf{Y}} \min_{y^* \in \mathbf{Y}^*} d(y, y^*), \max_{y^* \in \mathbf{Y}^*} \min_{y \in \mathbf{Y}} d(y, y^*)\right\}, \end{aligned} \tag{18}$$

$$J(\mathbf{Y}, \mathbf{Y}^*) = \frac{|\mathbf{Y} \cap \mathbf{Y}^*|}{|\mathbf{Y} \cup \mathbf{Y}^*|} = \frac{|\mathbf{Y} \cap \mathbf{Y}^*|}{|\mathbf{Y}| + |\mathbf{Y}^*| - |\mathbf{Y} \cup \mathbf{Y}^*|}, \tag{19}$$

$$Recall = \frac{TP}{TP + FN}, \tag{20}$$

$$Precision = \frac{TP}{TP + FP}. \tag{21}$$

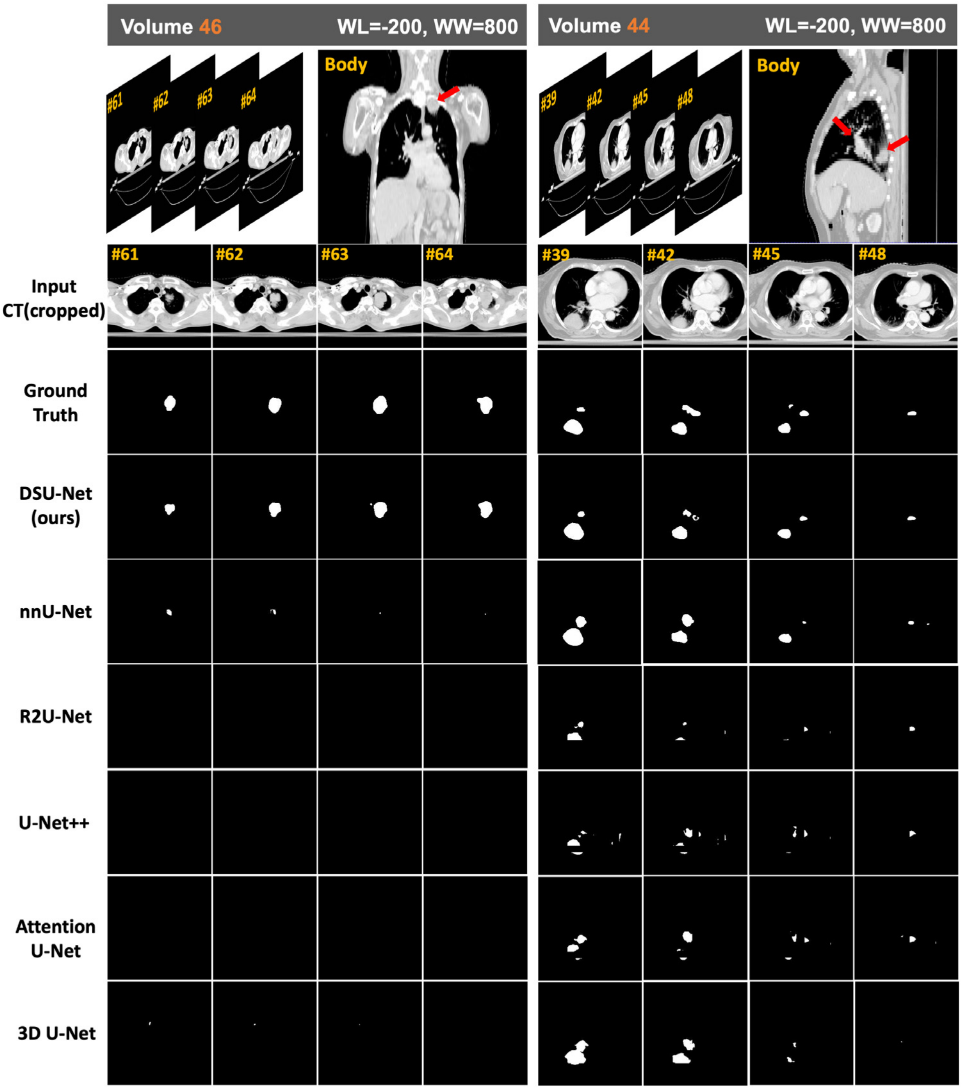

Fig. 4. Visual comparison on LC Dataset. Volume 46 and Volume 44 denote two patients CT volumes. WL and WW are short for window level and window width, respectively. Sag means sagittal plane and Cor means coronal plane. $\#X$ denotes the Xth slice in the transverse plane. For better illustration, Input CT (cropped) denotes the center cropped image. The two cases show that our DSU-Net can obtain better performance compared with other methods, especially in distraction areas.

4.2. Comparison with the state-of-the-art methods

In order to evaluate our model, we make comparison with five well-known U-like Nets, including 3D U-Net (Çiçek et al., 2016), nnU-Net (Isensee et al., 2018), Attention U-Net (Oktay et al., 2018), R2U-Net (Alom et al., 2018) and UNet++ (Zhou et al., 2018), and adopt their official code implementations. In order to make a fair comparison, we empirically finetuned hyperparameters of these models to get good results for each model. Then all the methods are tested on the LC testing dataset (as mentioned in Section 4.1).

4.2.1. Quantitative analysis

The quantitative comparison results are listed in Table 1. We can observe that the proposed DSU-Net has the competitive results of segmenting lung tumors. For the Dice metric, our DSU-Net gets 0.4455 which is the best result comparing to other U-like Nets. It achieves 9.76% higher Dice than the second highest nnU-Net(0.3479). For the 95% HD, 3D U-Net gets the best value 24.40, while DSU-Net achieves 32.2% better than the backbone nnU-Net. What is more, DSU-Net has the highest Jaccard value (0.3121), Recall value (0.3722), and Precision value (0.5990). On the other hand, nnU-Net gets the second highest for these metrics which has a lower 8.35% Jaccard, 4.58% Recall and 14.63%. The good results of our method are caused by the distraction information handling scheme in our network. Intuitively, our method increases the feature distinction by enhancing the positive region features and suppressing the negative region features.

Table 1
Comparing U-like Nets with our DSU-Net on the LC Dataset, we find that DSU-Net obtains the best results in terms of four evaluation indicators.

Method	Dice↑	95% HD↓	Jaccard↑	Recall↑	Precision↑
3D U-Net	0.2288	24.40	0.1423	0.1832	0.3755
Attention U-Net	0.1473	39.31	0.0902	0.1216	0.3760
R2U-Net	0.1769	28.19	0.1013	0.1471	0.4219
UNet++	0.23	30.06	0.1413	0.1841	0.4442
nnU-Net	0.3479	61.21	0.2286	0.3264	0.4527
DSU-Net	0.4455	41.49	0.3121	0.3722	0.5990

4.2.2. Visual comparison

As shown in Fig. 4, we report two cases (volume 46 and volume 44) to illustrate our experiment results. In order to better display

Table 2
Results of the ablation study. As shown in this table, **No. 1** is nnU-Net without considering distraction information. **No. 2** is a simplified version of DAM, which uses $\mathbf{F}_{fn}$ as the final output of FNAM. **No. 3** removes the adaptive pooling and extension operations in Fig. 3. **No. 4** is our final proposed DSU-Net.

No.	Dice↑	95%HD↓	Jaccard↑	Recall↑	Precision↑
1	0.3479	61.21	0.2286	0.3264	0.4527
2	0.2041	135.86	0.1243	0.2765	0.2011
3	0.4256	41.21	0.2786	0.3469	0.5931
4	0.4455	41.49	0.3121	0.3722	0.5990

the 3D lesions and the corresponding segmentation outputs, we show the lesions with a sequence of slices instead of a certain slice in the transverse plane. Besides, to better show the body location and the size of the tumor, we add the other projection planes (coronal plane for volume 46 and sagittal plane for volume 44). Here, in volume 46, four adjacent slices {#61, #62, #63, #64} are used to represent the tumor near the throat (as shown in the sagittal plane with a red arrow). In volume 44, there are two long strip tumors existing in the center of the chest, and we select four slices {#39, #42, #45, #48} to better observe the segmentation results in the whole tumor. To better visualize the tumor areas, we set the window level to −200 and the window width to 800. Moreover, since the tumors are too tiny relative to the whole slice, the center cropping operation is used to show the core chest areas (see Input CT(cropped)).

In volume 46, there is a spherical tumor near the throat with medium size. As observed by naked eyes, this tumor looks similar to the surrounding throat tissue. Most of the compared methods misdiagnose this tumor as normal tissue since it is hard to distinguish this tumor from the context information. In contrast, our DSU-Net can segment this tumor accurately even at the tumor boundaries. In volume 44, there are two long strip tumors in the center of the chest. Most compared methods can obtain the approximate location for the center of the tumor, while they may fail at the tumor boundaries to different certain degrees. On the other hand, our DSU-Net not only effectively locates the center of tumors but also handles well on the boundaries (just like slice #45 and slice #48).

4.3. Ablation study

To evaluate the performance of DSU-Net, we conduct experiments with different settings as shown in Table 2. (1) The first row (**No. 1**) is nnU-Net and there is no distraction information considered here.

(2) The second row (**No. 2**) is a simplified version of DAM. Here, we only reserve $\mathbf{F}_{fn}$ with its supervision, and use it as the final output of FNAM ($\mathbf{F}_{fne}$), while discarding other operations and branches. In this setting, we find that all the metrics declined with varying degrees compared with the first row (**No. 1**) (41% worse in Dice, 45% worse in Jaccard, 15% worse in Recall and 55% worse in Precision).

(3) The third row follows the FNAM structure while removing adaptive pooling and extension operations (Fig. 3). Compared to the second row (**No. 2**), the third row (**No. 3**) enhances some regions of original features via attention mechanism rather than using the directly learned distraction features. In this setting, all the evaluation metrics obtain the obvious improvements compared with the first row (**No. 1**) (22% better in Dice, 21% better in Jaccard, 6% better in Recall and 31% better in Precision).

(4) The fourth row is our final DSU-Net. Comparing with the third row (**No. 3**), the fourth row (**No. 4**) adds the adaptive pooling and extension operations. By these two operations, the module can help the distraction features to take some global information into account. The fourth row (**No. 4**) gets more improvement in performance compared with the first row (**No. 1**) (22% better in Dice, 36% better in Jaccard, 14% better in Recall and 32% better in Precision).

5. Conclusion

To address the challenges from the ambiguous regions of lung tumor and non-tumor tissues with similar visual appearances, we proposed the cascaded distraction sensitive U-Net. The distraction regions are learned from a global prediction, and used to guide the prediction of refined segmentation of tumor regions. The developed Distraction-Attention Module strengthens the information of those easily ignored parts, by enhancing the false negative aspects and suppressing the false positives. The experimental results demonstrate that the proposed DSU-Net achieves better performance in 3D segmentation of lung tumor, as compared with those state-of-the-art U-like methods.

CRediT authorship contribution statement

Junting Zhao: Conceptualization, Methodology, Writing – original draft, Investigation. **Meng Dang:** Methodology, Investigation, Experiment. **Zhihao Chen:** Methodology, Investigation, Experiment. **Liang Wan:** Supervision, 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.

Acknowledgment

This work was supported by Tianjin Municipal Natural Science Foundation (20JCYBJC00960).

References

Albu, A., Precup, R.-E., Teban, T.-A., 2019. Results and challenges of artificial neural networks used for decision-making and control in medical applications. Facta Universitatis, Series: Mechanical Engineering 17 (3), 285–308.

Alom, M.Z., Hasan, M., Yakopcic, C., Taha, T.M., Asari, V.K., 2018. Recurrent residual convolutional neural network based on u-net (R2U-net) for medical image segmentation. arXiv preprint arXiv:1802.06955.

Asuntha, A., Brindha, A., Indirani, S., Srinivasan, A., 2016. Lung cancer detection using SVM algorithm and optimization techniques. J. Chem. Pharm. Sci. 9 (4), 3198–3203.

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.

Borlea, I.-D., Precup, R.-E., Borlea, A.-B., Iercan, D., 2021. A unified form of fuzzy C-means and K-means algorithms and its partitional implementation. Knowl.-Based Syst. 214, 106731.

Cao, H., Liu, H., Song, E., Ma, G., Xu, X., Jin, R., Liu, T., Hung, C.-C., 2019. Two-stage convolutional neural network architecture for lung nodule detection. arXiv preprint arXiv:1905.03445.

Chan, T.F., Vese, L.A., 2001. Active contours without edges. IEEE Trans. Image Process. 10 (2), 266–277.

Chang, P., Teng, W., 2007. Exploiting the self-organizing map for medical image segmentation. In: Twentieth IEEE International Symposium on Computer-Based Medical Systems, CBMS'07, pp. 281–288.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 834–848.

Chen, S., Tan, X., Wang, B., Lu, H., Hu, X., Fu, Y., 2020. Reverse attention-based residual network for salient object detection. IEEE Trans. Image Process. 29, 3763–3776.

Chen, W., Wei, H., Peng, S., Sun, J., Qiao, X., Liu, B., 2019. HSN: hybrid segmentation network for small cell lung cancer segmentation. IEEE Access 7, 75591–75603.

Çiçek, O., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. arXiv preprint arXiv:1606.06650.

Dehmeshki, J., Amin, H., Valdivieso, M., Ye, X., 2008. Segmentation of pulmonary nodules in thoracic CT scans: A region growing approach. IEEE Trans. Med. Imaging 27 (4), 467–480.

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

Hu, Y., Grossberg, M., Mageras, G., 2008. TH-D-332-02: Semi-automatic medical image segmentation with adaptive local statistics in conditional random field framework. Med. Phys. 35 (6Part27).

Huang, Q., Wu, C., Xia, C., Wang, Y., Kuo, C.-C.J., 2017. Semantic segmentation with reverse attention. In: Proceedings of the British Machine Vision Conference. BMVC, BMVA Press, pp. 18.1–18.13.

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al., 2018. Nnu-net: Self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486.

Kostis, W.J., Reeves, A.P., Yankelevitz, D.F., Henschke, C.I., 2003. Three-dimensional segmentation and growth-rate estimation of small pulmonary nodules in helical CT images. IEEE Trans. Med. Imaging 22 (10), 1259–1274.

Li, L., Zhao, X., Lu, W., Tan, S., 2020. Deep learning for variational multimodality tumor segmentation in PET/CT. Neurocomputing 392, 277–295.

Liu, X., Jiang, T., Li, W., Li, X., Zhao, C., Shi, J., Zhao, S., Jia, Y., Qiao, M., Zhang, L., Luo, J., Gao, G., Zhou, F., Wu, F., Chen, X., He, Y., Ren, S., Su, C., Zhou, C., 2018. Characterization of never-smoking and its association with clinical outcomes in Chinese patients with small-cell lung cancer. Lung Cancer 115, 109–115.

Liu, T., Tian, Y., Zhao, S., Huang, X., Wang, Q., 2019. Automatic whole heart segmentation using a two-stage U-net framework and an adaptive threshold window. IEEE Access 7, 83628–83636.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Mishra, M., Nayak, J., Naik, B., Abraham, A., 2020. Deep learning in electrical utility industry: a comprehensive review of a decade of research. Eng. Appl. Artif. Intell. 96, 104000.

Nayak, J., Naik, B., Behera, H.S., Abraham, A., 2018. Elitist teaching–learning-based optimization (ETLBO) with higher-order Jordan pi-sigma neural network: a comparative performance analysis. Neural Comput. Appl. 30 (5), 1445–1468.

Netto, S.M.B., Silva, A.C., Nunes, R.A., Gattass, M., 2012. Automatic segmentation of lung nodules with growing neural gas and support vector machine. Comput. Biol. Med. 42 (11), 1110–1121.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al., 2018. Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

Pan, Z., Lu, J., 2007. A Bayes-based region-growing algorithm for medical image segmentation. Comput. Ence Eng. 9 (4), 32–38.

Pozna, C., Precup, R.-E., 2014. Applications of signatures to expert systems modelling. Acta Polytech. Hung. 11 (2), 21–39.

Pratiksha, H., 2017. Texture based interstitial lung disease detection using convolutional neural network. In: 2017 International Conference on Big Data, IoT and Data Science. BID, IEEE, pp. 18–22.

Reboucas, P.P., Cortez, P.C., Holanda, M.A., 2011. Active contour modes crisp: new technique for segmentation of the lungs in CT images. Rev. Bras. de Eng. Bioméd. 27 (4), 259–272.

Reboucas Filho, P.P., Cortez, P.C., da Silva Barros, A.C., Albuquerque, V.H.C., Tavares, J.M.R., 2017. Novel and powerful 3D adaptive crisp active contour method applied in the segmentation of CT lung images. Med. Image Anal. 35, 503–516.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Roth, H.R., Oda, H., Zhou, X., Shimizu, N., Yang, Y., Hayashi, Y., Oda, M., Fujiwara, M., Misawa, K., Mori, K., 2018. An application of cascaded 3D fully convolutional networks for medical image segmentation. Comput. Med. Imaging Graph. 66, 90–99.

Sato, M., Lakare, S., Wan, M., Kaufman, A., Nakajima, M., 2002. A gradient magnitude based region growing algorithm for accurate segmentation. In: International Conference on Image Processing.

Shen, D., Wu, G., Suk, H.-I., 2017. Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. 19 (1), 221–248.

Sujji, G.E., Lakshmi, Y.V.S., Jiji, G.W., 2013. MRI brain image segmentation based on thresholding. Int. J. Adv. Comput. Res. 3 (1), 97–101.

Suster, D.I., Mino-Kenudson, M., 2020. Molecular pathology of primary non-small cell lung cancer. Arch. Med. Res..

Taheri, S., Ong, S.H., Chong, V.F.H., 2010. Level-set segmentation of brain tumors using a threshold-based speed function. Image Vis. Comput. 28 (1), 26–37.

Wang, C., Xu, R., Xu, S., Meng, W., Xiao, J., Peng, Q., Zhang, X., 2020. Accurate 2D soft segmentation of medical image via SoftGAN network. arXiv preprint arXiv:2007.14556.

Xiao, H., Feng, J., Wei, Y., Zhang, M., Yan, S., 2018. Deep salient object detection with dense connections and distraction diagnosis. IEEE Trans. Multimed. 20 (12), 3239–3251.

Xu, R., Pan, J., Hirano, Y., Ye, X., Kido, S., Tanaka, S., 2017. A pilot study to utilize a deep convolutional network to segment lungs with complex opacities. In: 2017 Chinese Automation Congress. CAC, IEEE, pp. 3291–3295.

Yazdanpanah, A., Hamarneh, G., Smith, B., Sarunic, M., 2009. Intra-retinal layer segmentation in optical coherence tomography using an active contour approach. In: Medical Image Computing & Computer-Assisted Intervention.

Zall, R., Kangavari, M.R., 2019. On the construction of multi-relational classifier based on canonical correlation analysis. Int. J. Artif. Intell. 17 (2), 23–43.

Zhao, B., Cao, Z., Wang, S., 2017. Lung vessel segmentation based on random forests. Electron. Lett. 53 (4), 220–222.

Zhao, N., Tong, N., Ruan, D., Sheng, K., 2019. Fully automated pancreas segmentation with two-stage 3d convolutional neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 201–209.

Zheng, Q., Qiao, X., Cao, Y., Lau, R.W., 2019. Distraction-aware shadow detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5167–5176.

Zhong, Z., Kim, Y., Zhou, L., Plichta, K., Allen, B., Buatti, J., Wu, X., 2018. 3D fully convolutional networks for co-segmentation of tumors on PET-ct images. In: 2018 IEEE 15th International Symposium on Biomedical Imaging. ISBI 2018, pp. 228–231.

Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 3–11.

Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., Hu, W., 2018. Distractor-aware siamese networks for visual object tracking. In: Proceedings of the European Conference on Computer Vision, ECCV.