Computerized Medical Imaging and Graphics 92 (2021) 101957

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

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

MSDS-UNet: A multi-scale deeply supervised 3D U-Net for automatic segmentation of lung tumor in CT

Jinzhu Yang [a,b], Bo Wu [a,b], Lanting Li [a,b], Peng Cao [a,b,*], Osmar Zaiane [c]

[a] Computer Science and Engineering, Northeastern University, Shenyang, China
[b] Key Laboratory of Intelligent Computing in Medical Image of Ministry of Education, Northeastern University, Shenyang, China
[c] Amii, University of Alberta, Edmonton, Alberta, Canada

ARTICLE INFO

Keywords:
Lung tumors segmentation
3D U-Net
Deep supervision
Multi-scale tumors

ABSTRACT

Lung cancer is one of the most common and deadly malignant cancers. Accurate lung tumor segmentation from CT is therefore very important for correct diagnosis and treatment planning. The automated lung tumor segmentation is challenging due to the high variance in appearance and shape of the targeting tumors. To overcome the challenge, we present an effective 3D U-Net equipped with ResNet architecture and a two-pathway deep supervision mechanism to increase the network's capacity for learning richer representations of lung tumors from global and local perspectives.

Extensive experiments on two real medical datasets: the lung CT dataset from Liaoning Cancer Hospital in China with 220 cases and the public dataset of TCIA with 422 cases. Our experiments demonstrate that our model achieves an average dice score (0.675), sensitivity (0.731) and *F*1-score (0.682) on the dataset from Liaoning Cancer Hospital, and an average dice score (0.691), sensitivity (0.746) and *F*1-score (0.724) on the TCIA dataset, respectively. The results demonstrate that the proposed 3D MSDS-UNet outperforms the state-of-the-art segmentation models for segmenting all scales of tumors, especially for small tumors. Moreover, we evaluated our proposed MSDS-UNet on another challenging volumetric medical image segmentation task: COVID-19 lung infection segmentation, which shows consistent improvement in the segmentation performance.

1. Introduction

Lung cancer is one of the major public health issues that seriously threatens the health of humans (Hoffman et al., 2000). Lung cancer has a high mortality rate, especially advanced lung cancer, which is difficult to cure and more likely to metastasize. Computed tomography (CT) is widely used for computer aided diagnosis of lung cancer. Fig. 1 shows some examples of lung tumors in CT. The segmentation of lung tumor is a topic of great interest in medical image analysis since it provides doctors with meaningful and reliable quantitative information in diagnosing and monitoring neurological diseases.

For a better understanding of the pathophysiology of cancer, quantitative imaging can reveal clues about the disease characteristics and effects on particular anatomical structures. Segmentation and the subsequent quantitative assessment of lung tumors in medical images provide valuable information for the analysis of pathologies and are important for planning of treatment strategies, monitoring of disease progression and prediction of patient outcome. However automated lung tumor segmentation is challenging due to the high variance in appearance and shape of the targeting tumors (Kamal et al., 2018; Hossain et al., 2019; Yang et al., 2018; Jiang et al., 2018b). The heterogeneous appearance of lesions including the large variability in location, size and shape between the patients make it difficult to devise effective segmentation rules. Moreover, there was a wide variation in distribution of tumors across populations and datasets. The variation of tumors are shown in Fig. 2. As observed in Fig. 2, the segmentation problem is quite challenging and the difficulties. The traditional segmentation methods are based on hand-crafted or shallow-learning-based features with limited representation power, resulting in failing to provide strong representation capability to deal with the large variations of tumor appearance and shape. Recently, deep learning algorithms, especially U-Net (Çiçek et al., 2016), has shown their much stronger detection power in biomedical image segmentation applications (Jiang et al., 2018a; Long et al., 2015). Deep learning methods for segmentation can automatically learn hierarchies of relevant features directly from the training data in an end-to-end manner. The U-Net network

* Corresponding author at: Computer Science and Engineering, Northeastern University, Shenyang, China.
E-mail address: caopeng@mail.neu.edu.cn (P. Cao).

https://doi.org/10.1016/j.compmedimag.2021.101957
Received 28 October 2020; Received in revised form 5 March 2021; Accepted 8 July 2021
Available online 24 July 2021
0895-6111/© 2021 Published by Elsevier Ltd.

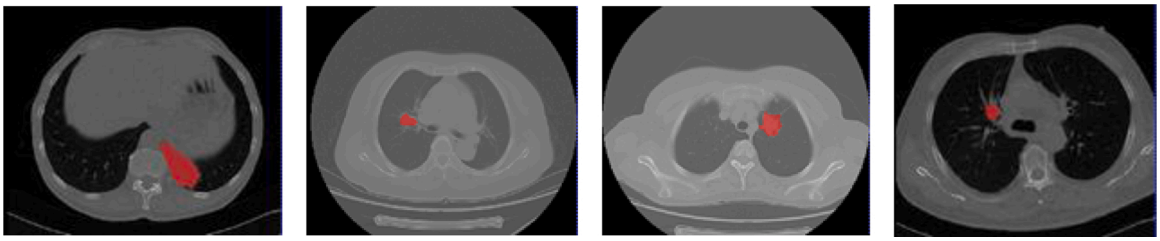


Fig. 1. Some examples of lung tumors. The red part indicates the lung tumors.

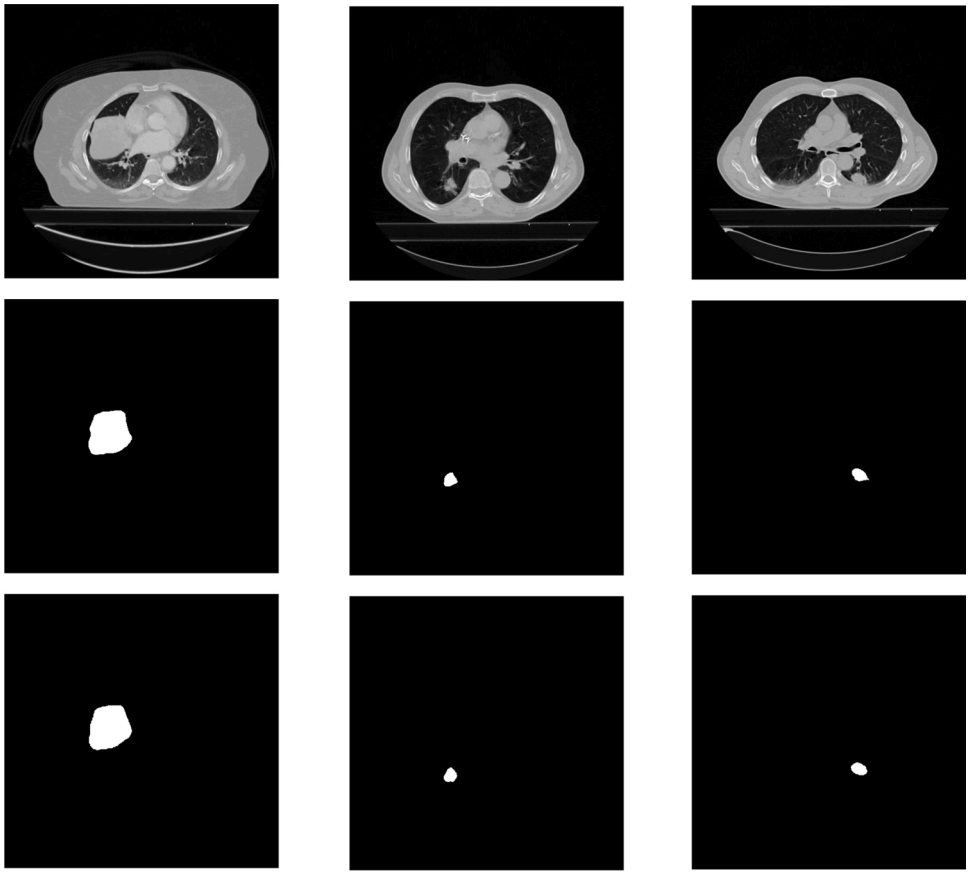
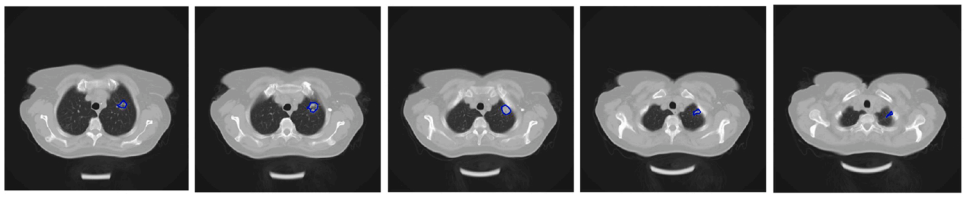


Fig. 2. The variation of lung tumors and the comparison between ground truth and binary masks predicted by our proposed methods. The 1–3 rows indicates the original lung image, ground truth and the predicted masks by our proposed method. Moreover, (a) is a large tumor, (b) is a small tumor and (c) is a tumor attached to the pleura, all of which are obtained from a local hospital in China.

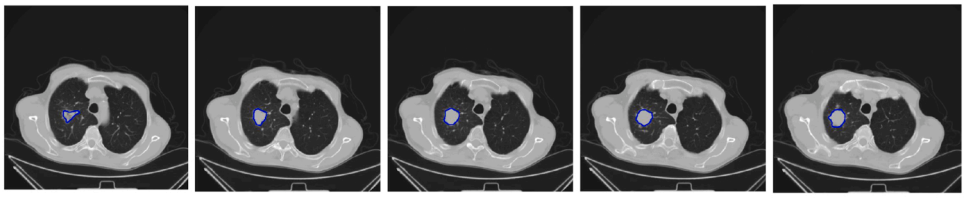
consists of an encoder performing feature extraction and a decoder performing information fusion (Li et al., 2018; Badrinarayanan et al., 2017). U-Net adopts the skip connection for feature fusion to achieve multi-level resolution information utilization and avoid feature loss (Ronneberger et al., 2015; Chen and Qi, 2018). For the U-Net based deep learning methods, tumor segmentation can be regarded as a pixel-wise binary classification problem, in which each pixel is classified as a tumor pixel or non-tumor pixel (Chen and Qi, 2018; Guofeng et al., 2018; Ibtehaz and Rahman, 2019).

The U-Net based models have proven their effectiveness and superiority over traditional medical image segmentation algorithms (Huang et al., 2018; Li et al., 2019). However, they are conceivably not optimal for volumetric medical image analysis as they cannot take full advantage of the special information encoded in the volumetric data. The 2D U-Net model for segmenting lung tumors only obtain a single tumor slice in CT images, while lung tumors are usually distributed in continuous CT slices. The volumetric medical image segmentation is a fundamental yet challenging problem in medical image analysis. Several 3D volume-to-volume segmentation networks have been proposed, including 3D U-Net (Çiçek et al., 2016), V-Net (Milletari et al., 2016) and 3D CNN (Dou et al., 2017a). Although these 3D segmentation networks can improve the volumetric tumor segmentation by capturing representative features across all three spatial dimensions, they still present limited capacity in effectively learning the feature information of the images in complicated tasks, such as the segmentation of heterogeneous lung tumors. The reasons that limits the learning performance of models are:

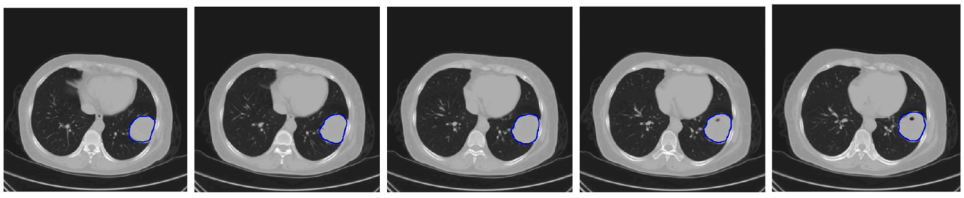
1 The segmentation of lung tumor is typically a difficult task due to the large heterogeneity of cancer lesions. The different subtypes of cell



(a) Small scale



(b) Medium scale



(c) Large scale

Fig. 3. Multi-scale lung tumors.

carcinoma can bring diverse intensity attribution and scales in CT images. Fig. 3 shows lung CT images from three different patients. Because the location and scale of tumor vary considerable across patients, the segmentation of lung tumor is challenging. In particular, the largest tumor occupies over one million voxels, but the smallest one has only thousands. This motivates us to train multi-scale networks to deal with such a large variation in scale (Fan et al., 2020b; Xu et al., 2018; Ma et al., 2018; Kamnitsas et al., 2017). Therefore, a network should be robust enough to analyze objects at different scales. The previous work usually uses Atrous spatial pyramid pooling (ASPP) to robustly segment objects at multiple scales (Chen et al., 2017), or captures the multi-scale information by stacking multi-scale images as inputs. The limitations are that they result in more parameters to be trained and the multi-scale representation is learned independently with ignoring the correlation of multiple scale learning. Furthermore, most previous studies fail to segment small tumors, such as dilated convolution, does not work well with small tumor, which may have a significant impact on finding early-stage cancers.

2 The standard U-Net architecture contains only a few layers and therefore it is not currently deep enough to gain improved performance over other existing networks. To solve the problem, adding more layers directly to the network can enlarge the parameter space and make the network deeper, which may lead to gradient vanishing and redundant computation during training. Gradient vanishing means if the network contains too many hidden layers, the learning rate will decrease with forward propagation, which may decrease the overall network learning.

3 The 3D medical images have much more complicated anatomical environments than 2D images, hence 3D variants of U-Net with much more parameters are usually required to capture more representative features. However, the extensive number of parameter weights and depth in 3D U-Net introduces various optimization difficulties, such as over-fitting and slow convergence speed.

4 Training such a 3D networks often confronts various optimization difficulties, and the insufficiency of training samples (patients) hinders the training of the segmentation model. It is more difficult for 3D cases than 2D when only a small set of annotated data is available.

To solve the issues, we extend the traditional 2D U-Net to a 3D version equipped with ResNet architecture to capture the inter-slice continuity of the tumor as a solid object in our work, and propose a multi-scale multi-level deep supervision U-Net (MSDS-UNet) that integrates the existing 3D U-Net model with a two-pathway deep supervision mechanism for more accurate segmentation performances. Moreover, the ResNet module (He et al., 2016) is adopted in each block to increase the network's capacity for learning richer representations. Deep supervision was firstly introduced by Lee et al. (2015) as a way to deal with the problem of the vanishing gradient in training deeper CNN for image classification. In the literature, some segmentation methods with the scheme of deep supervision have been developed and studied. In the case of medical applications, it has been employed to prostate segmentation (Zhu et al., 2017), liver (Chung et al., 2020), and kidney tumor (Tureckova et al., 2020) segmentation in CT volumes, and to brain tumor segmentation from magnetic resonance imaging (Isensee et al., 2017). Moreover, the deep supervision is incorporated with multiple network models, such as CNN (Dou et al., 2017b) and U-Net (Zeng et al., 2017). These methods aim to improve the convergence rate and discrimination capability of deep learning models. We evaluated MSDS-UNet using three medical imaging datasets covering lung tumor segmentation and COVID-19 infection segmentation. The experimental results have indicated the effectiveness of the proposed improvements and suggest that our approach could acquire competitive performance as state-of-the-art multi-scale lung tumor or infected lesion segmentation methods. The contribution of our work is as follow:

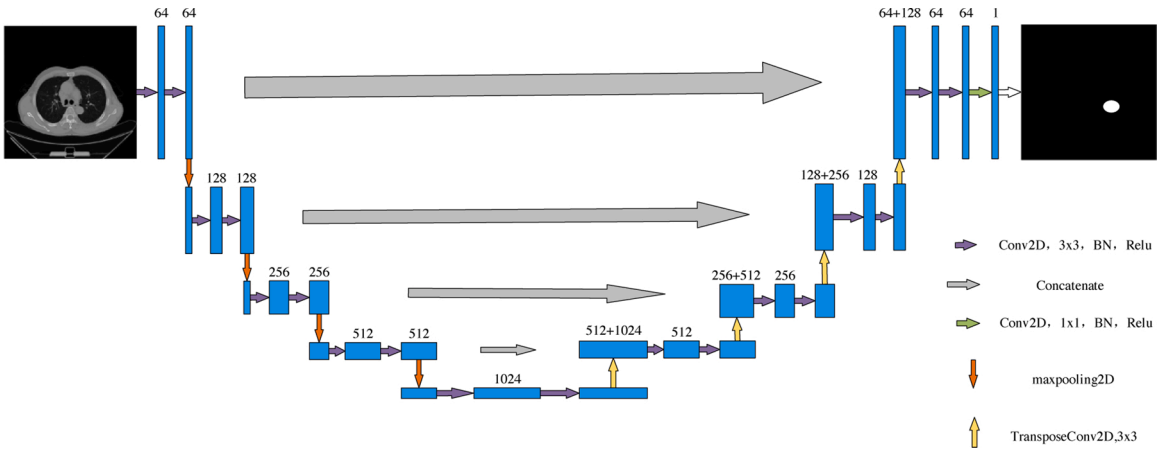


Fig. 4. 2D U-Net framework.

(1) We develop a volume-to-volume tumor segmentation network to automatically segment lung tumors from CT images. The framework integrates the existing 3D U-Net model with a 3D deep supervision mechanism to capture the inter-slice continuity of the tumor and achieve more accurate segmentation performances.

(2) We further propose a 3D deep supervision mechanism by formulating an objective function that directly guides the training of the hidden layers in order to reinforce the propagation of gradient flows within the network and hence learn more powerful and representative features. The previous work does not take full advantage of the deep supervision mechanism. We propose a two-pathway deep supervision in U-Net, improving the segmentation performance from two aspects: (1) multiple predictions from multiple semantic layers are generated and averaged to produce an accurate segmentation with the help of deep supervision; (2) regularizing the weights of layers with local deep supervision for the learned features. Although the local deep supervision does not work in the inference stage, it further improves the segmentation through the proper regularization of the network weights. Both of them improve the performance, as we show empirically in Section 4.1.2.

(3) As pointed out in most previous works, a good tumor segmentation network should be deep enough such that multi-level features can be learned. It should have multiple stages to learn more inherent features from different scales. In this paper, we focus on the deep supervision instead of simply fusing the multi-level features extracted from different scales. Different the popular methods with fusing low-level but high-resolution features and high-level low-resolution features together, our two-pathway deeply supervised U-Net can improve low-level features and mid-level features by assigning auxiliary supervisions directly to the early stages of the network. To generate discriminative outputs in the auxiliary branches, low-level and mid-level features are forced to encode more semantic concepts, which is expected to be helpful for the final segmentation. Moreover, different from previous deeply supervision architectures, our model is a two-pathway deep supervision involving supervisions for hard and soft fusion of the side-output predictions for the final prediction, and supervisions for the ensemble of multi-scale predictions of lower resolution segmentation map, to increase the network's capacity for learning richer representations of lung tumors. In our work, we present a comprehensive analysis to better understand the representations learned with the help of deep supervision can derive better representations of lung tumor at different scales. Our experimental results yield a solid evidence that imposing a deeply supervised method during training the network is a viable method for improving U-Net's segmentation performances for lesions that appear at multiple scales. In particular, the traditional U-Net does not work well on small tumors, but MSDS-UNet shows obvious improvement.

(4) The performance of deep supervision is highly dependent on an appropriate choice of weights among all losses of all tasks. In other words, different tasks need to be properly balanced, so that the model can converge to the state which are useful across all tasks. Unlike the previous work, we consider it as a multi-task learning formulation and assign different weights for different tasks. In order to automatically achieve an optimal weights for multi-task learning and reduce the chance that negative transfer happens, we utilize GradNorm (Chen et al., 2018) to learn a balanced global task weight. The scheme is able to avoid certain group of related tasks dominates the training process and the tasks outside the dominant group cannot be optimized sufficiently.

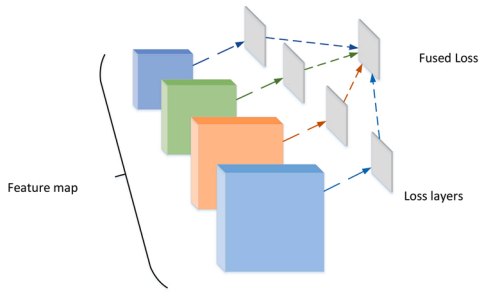
(5) Other than improving the state-of-the-art results, we conduct exhaustive analysis on the behavior of deep supervision for 3D U-Net on the multi-scale lung tumors systematically, especially on the small scale cancers.

The rest of the paper is organized as follows. Section 2 introduce 2D U-Net. In Section 3 we describe the architecture of our network and the procedure of the proposed deep supervision. Section 4 presents and discuss the experimental results. Section 5 discusses the limitation and the future research. At last, this paper is concluded in Section 6.

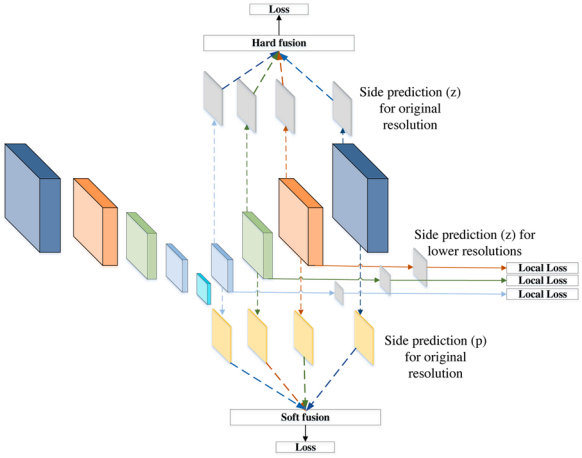
2. U-Net framework

U-Net is an U-shaped convolutional neural network used for image segmentation (Norman et al., 2018). Fig. 4 shows the framework of the 2D U-Net. The network architecture is symmetric, having an Encoder that extracts spatial features from the image, and a Decoder that constructs the segmentation map from the encoded features. The Encoder follows the typical formation of a convolutional network.

The encoder involves a sequence of two 3×3 convolution operations, followed by a max-pooling operation with a pooling size of 2×2 and stride of 2. This sequence is repeated four times, finally, a progression of two 3×3 convolution operations connects the Encoder to the Decoder. Similar to the encoding phase, the decoder replaces the pooling with the up-sampling using a 2×2 transposed convolution operation and connects the corresponding feature maps in the two stage to complete the information fusion. Similar to the encoder, a series of up-sampling and two 3×3 convolution operations are repeated four times in the decoder part, halving the number of filters at each stage. Finally, a 1×1 convolution operation is performed to generate the final segmentation map. A Sigmoid is used as the activation function to



(a) The original deep supervision



(b) The proposed two-pathway deep supervision

Fig. 5. The illustration of original deep supervision and our proposed two-pathway deep supervision.

normalize the output value between 0 and 1 which represents the probability that each pixel belongs to a tumor. In both parts, the batch normalization layer is involved to normalize the output features and the ReLu activation function is used to increase the network's nonlinear expression ability. The smallest feature map size in the encoding stage is $8 \times 8 \times 8$.

3. Methods

3.1. The network architecture of 3D MSDS-UNet

The scale and appearance of the tumor often vary obviously among patients, which is a great challenge when training deep models for lung tumor segmentation. Inspired by deeply-supervised networks applied to image classification and segmentation, we introduce a new deep supervision mechanism into our 3D U-Net model to deal with the problems noted above. Different from previous deeply supervision architectures, a combination of direct multi-scale side prediction and multi-level segmentation fusion are proposed, to increase the network's capacity for learning richer representations of lung tumors. The Fig. 5 illustrates the original deep supervision and our proposed deeply-supervision pathways strategy.

In order to achieve efficient end-to-end learning and inference, we first develop a 3D U-Net network for volumetric image segmentation. To improve the representation capacity of the segmentation network and to optimize the segmentation performance, we modify the 3D U-Net architecture with a ResNet module (He et al., 2016), which has been experimentally proven to enhance the capturing of more visual information. The ResNet module is adopted in each block to increase the network's capacity for learning richer representations. The residual networks are easy to optimize can gain accuracy from considerably increased depth with avoiding the degradation problem. We further propose a 3D deep supervision mechanism by formulating a multi-level multi-scale objective function that directly guides the training of hidden layers, which can accelerate the convergence speed and improve the segmentation performance of the network. Fig. 6 illustrates the architecture of our proposed 3D MSDS-UNet network. Our network also consists of two parts: the encoder part focusing on the analysis and feature representation learning, and the decoder part generating segmentation results relying on the learned features from the encoder.

We employ parallel pathways for deep supervision in 3D UNet, a solution to effectively incorporate both local and global supervision information which greatly improves segmentation results.

3.2. 3D convolution and pooling

The 3D U-Net is accomplished by replacing the 2D layers with 3D layers, and we choose it as our base model to achieve the lung tumor segmentation accurately. 3D U-Net is formed by 3D convolution, 3D pooling, 3D batch normalization and activation function on the basis of the 2D U-Net, which can maximally take advantage of the spatial information (Christ et al., 2016; Kamal et al., 2018). These 3D operations are illustrated in Fig. 7.

3.3. The multi-level multi-scale deep supervision

In order to improve the extracting abilities of the 3D U-Net, we adopt deep supervision (Albarazanchi et al., 2016; Zeiler et al., 2010). In the deep supervised nets, the extra companion objectives are introduced to the individual hidden layers, in addition to the overall objective at the output layer. The segmentation performance of small tumors depend on the underlying features contained in the shallow hidden layer, while the identification of large tumor depends on the advanced features. Therefore, the side outputs of each hidden layers are predicted in each stage of the 3D U-Net with ResNet. For the coarse layer, it generates segmentation results with the coarsest resolution, while the output at the middle and the fine scales generate segmentation results with the intermediate and the finest resolutions, respectively. In order to achieve a multi-scale tumors segmentation, the deep supervision mechanism is brought in by associating a companion local output with each hidden layer from local and global perspectives.

Specifically, $\boldsymbol{D} = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}, i = 1, \ldots, N$, $\boldsymbol{x}_i$ is the input volumetric CT image, $\boldsymbol{y}_i$ denotes the corresponding ground truth map for $\boldsymbol{x}_i$. Moreover, let $\boldsymbol{W}$ be the weights of the main network, $\boldsymbol{w} = \{\boldsymbol{w}_0, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_M\}$ and $\boldsymbol{w}^p = \{\boldsymbol{w}_0^p, \boldsymbol{w}_1^p, \ldots, \boldsymbol{w}_M^p\}$ be the global fused weight and local weight parameters of side outputs at different scales, where $\boldsymbol{w}_m$ and $\boldsymbol{w}_m^p$ are the weights of the side models at scale of m, M is the number of sides and $M = 4$ in our study.

The local side output is to predict the segmentation of lower scales prediction in each layer by directly connecting to a 3D convolution and a one-channel convolutional layer with the kernel size 1×1. It predict the multi-level segmentation results with the same scales as the corresponding feature map. The ground truth is down-sampled to the same resolution as the feature map at first. It supervises the model to hierarchically learn the tumor segmentation with different scale of ground truth by taking advantages of multi-scale tumor context. These auxiliary losses of multi-scale local side prediction are proposed to hierarchically segment the lung tumors with multi-scale context. By making appropriate use of deeply supervising at each hidden layer of the network, we are able to directly influence the hidden layer weight update process to favor highly discriminative feature maps for segmentation.

The dice loss function is chosen as the loss function for supervising each side output:

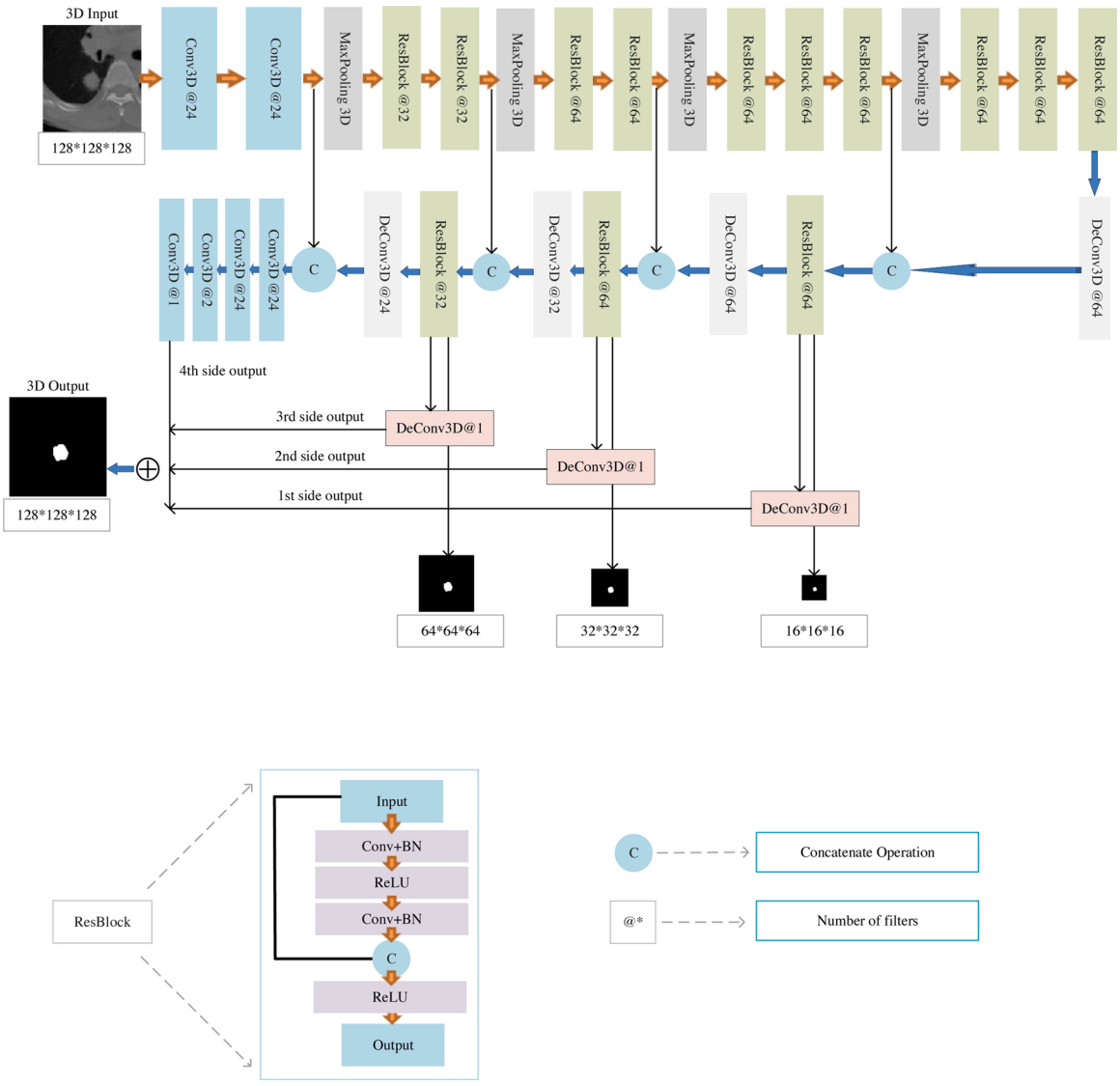


Fig. 6. 3D MSDS-UNet network structure. We employ parallel pathways for deep supervision in 3D UNet, a solution to effectively incorporate both local and global supervision information which greatly improves segmentation results. For obtaining better representation of lower layers, supervision is directly fed into corresponding layers. Besides, we add another convolutional layers or deconvolutional layers sizes in each side output. Such deep supervision learning strategy boosts the performance via: (1) directly constructing multi-level representations with multi-scale context; and (2) improving discrimination of intermediate layers, thus gaining improvement of overall performance.

$$f_{\text{dice}}\left(\boldsymbol{y}_m, \boldsymbol{z}_m^{(\boldsymbol{W},\boldsymbol{w}_m^p)}\right) = \frac{\left|\boldsymbol{y}_m \cap \boldsymbol{z}_m^{(\boldsymbol{W},\boldsymbol{w}_m^p)}\right|}{\left|\boldsymbol{y}_m \cap \boldsymbol{z}_m^{(\boldsymbol{W},\boldsymbol{w}_m^p)}\right| + 0.5\left(\left|\boldsymbol{y}_m - \boldsymbol{z}_m^{(\boldsymbol{W},\boldsymbol{w}_m^p)}\right| + \left|\boldsymbol{z}_m^{(\boldsymbol{W},\boldsymbol{w}_m^p)} - \boldsymbol{y}_m\right|\right)} \tag{1}$$

where $\boldsymbol{y}_m$ is the down-scaled segmentation map of the ground truth, the resulting segmentation map with corresponding scale is $\boldsymbol{z}_m^{(\boldsymbol{W},\boldsymbol{w}_m^p)}$, $\left|\boldsymbol{z}_m \cap \boldsymbol{y}_m\right|$ represents the number of correctly classified tumor voxels. $\left|\boldsymbol{z}_m^{(\boldsymbol{W},\boldsymbol{w}_m^p)} - \boldsymbol{y}_m\right|$ and $\left|\boldsymbol{y}_m - \boldsymbol{z}_m^{(\boldsymbol{W},\boldsymbol{w}_m^p)}\right|$ represents the number of all misclassified voxels. i.e. false positive or false negative.

Based on the dice loss function, the local loss for deep supervision at the hidden layers can be expressed as:

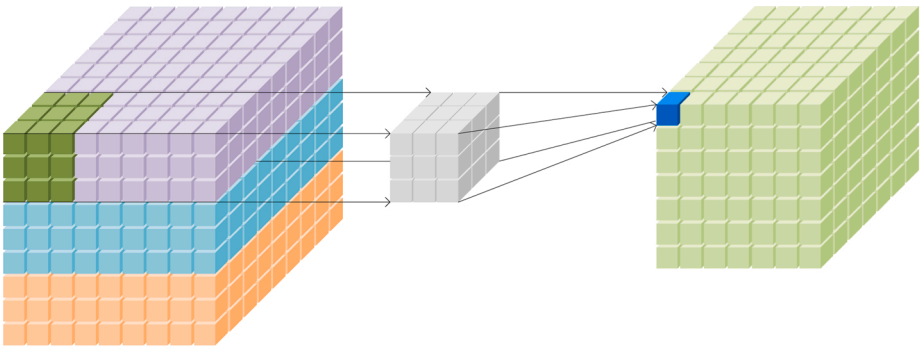
$$L_{\text{local}}(\boldsymbol{W}, \boldsymbol{w}^p) = \sum_{m=1}^{M} \alpha_m f_{\text{dice}}\left(\boldsymbol{y}_m, \boldsymbol{z}_m^{(\boldsymbol{W},\boldsymbol{w}_m^p)}\right), \tag{2}$$

where α_m is the weight of the mth side loss.

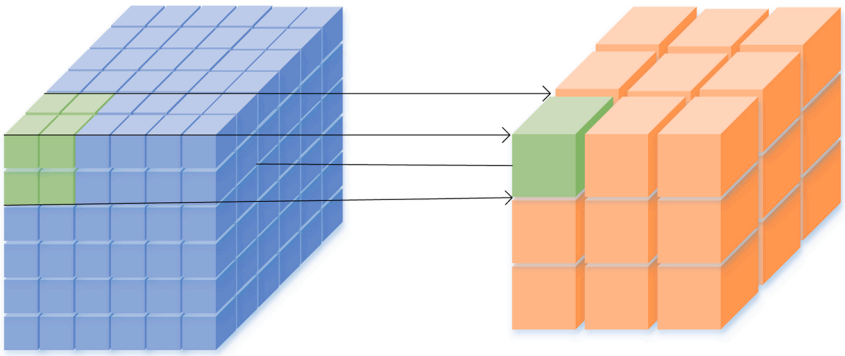
Therefore, in addition to the prediction of the main network, segmentation is performed at multiple output layers. The side at the coarse scale generates segmentation results with the coarsest resolution, while the models at the middle and the fine scales generate segmentation results with the intermediate and the finest resolutions, respectively. Then the abilities of feature representation are strengthened by minimizing the loss function between the local output segmentation map of each layer and the ground truth with corresponding scale.

Other than a series of direct local side losses are added after each side output for predicting the different scale segmentation, the multiple side predictions are fused into our network to generate the segmentation result with the original scale by up-scaling the segmentation maps of lower scales from global perspective. Therefore, the other output of each hidden layer is directly connected to a one-channel convolutional layer with the kernel size 1×1 followed by an 3D deconvolutional layer. The deconvolutional blocks are injected into lower layers such that the low-level and middle-level features are up-scaled to generate segmentation predictions with the same resolution as the input data.

Hard fusion: The hard fusion loss at the fusion layer can be expressed as:



(a) 3D Convolution: It represents a 3D convolution process of a feature map with a size of $9 \times 9 \times 9$ and a $3 \times 3 \times 3$ convolution kernel.



(b) 3D Pooling: It represents a 3D pooling process of a feature map with a $2 \times 2 \times 2$ sliding window.

Fig. 7. 3D convolution and pooling.

$$L_{\text{hard}}(\boldsymbol{W}, \boldsymbol{w}) = \sum_{m=1}^{M} \alpha_m f_{\text{dice}}(\boldsymbol{y}, \boldsymbol{z}^{(\boldsymbol{W}, \boldsymbol{w}^{(m)})}), \tag{3}$$

where α_m is the weight of the mth side loss.

Soft fusion: The soft fusion loss at the fusion layer can be expressed as:

$$L_{\text{soft}}(\boldsymbol{W}, \boldsymbol{w}) = f_{\text{dice}}(\boldsymbol{y}, g(\sum_{m=1}^{M} \alpha_m \boldsymbol{p}^{(m)})) \tag{4}$$

where $\boldsymbol{p}^{(m)}$ is the activations of the mth side output, where each value indicates the probability prediction $q^{(m)}(z_i|v_i; \boldsymbol{W}, \boldsymbol{w}^{(m)})$ of each voxel v_i, α_m is the weight of the mth side loss, $g(\cdot)$ denotes the sigmoid function.

Therefore, based on the multiple level supervisions in the hidden layers, a global loss is obtained by

$$L = L_{\text{hard}}(\boldsymbol{W}, \boldsymbol{w}) + \lambda_1 L_{\text{soft}}(\boldsymbol{W}, \boldsymbol{w}) + \lambda_2 L_{\text{local}}(\boldsymbol{W}, \boldsymbol{w}^p) + \lambda_3(\|\boldsymbol{W}\|_2^2 + \|\boldsymbol{w}\|_2^2 + \|\boldsymbol{w}^p\|_2^2) \tag{5}$$

where λ_1, λ_2 and λ_3 are all positive parameters which control contributions of hard fusion loss, soft fusion loss, local side loss and regularization, respectively. To prevent overfitting during model training, the weights are constrained using L2 regularization.

These auxiliary losses together with the loss from the last output layer are integrated to energize the back-propagation of gradients for more effective parameter updating in each iteration. At last, an ensemble mode is designed, where the segmentation results from all global segmentation branches (hard and soft fusion) are collected and then averaged. The multi-scale local side predictions only work in the training phrase to further help optimize the parameters in the main network. They work as regularization to the optimization of the network parameters.

The performance of our model is highly dependent on an appropriate choice of weights α among all losses of all tasks. In other words, different tasks need to be properly balanced, so that the model can converge to the state which are useful across all tasks. The challenge is to find the best value for each task at each training step that balances the contribution of each task for optimal model training. We incorporated gradient normalization (GradNorm) algorithm (Chen et al., 2018) into the optimization of our model, which enables automatically balance training in deep multi-task models by dynamically tuning gradient magnitudes. The gradient normalization can indicate whether the certain task is well trained or not, and decreases the relative weights of the well trained tasks. It increases the weight of a given task's loss when learning on that task is slower than other tasks. It has been demonstrated that it is effective for reducing negative transfer (Liu et al., 2019). The learned parameters are divide into two parts: the shared parameters $\boldsymbol{W}_s$ (the common parameters among the multiple tasks) and the specific parameter $\boldsymbol{W}_m$ of each task involving the parameters in certain layers of the main network and the parameters $\boldsymbol{w}$ and $\boldsymbol{w}^p$ in the corresponding layer. GradNorm is applied on the optimization of the shared parameters $\boldsymbol{W}_s$. In additional, each task loss L_m involves the hard fusion loss and the local loss. The reason is that the soft fusion loss cannot explicitly decomposed into multiple separable tasks. The procedure of the network optimization with GradNorm is shown in Algorithm 1.

Table 1
The segmentation results of tumors with different U-Net based segmentation models. Superscript symbols * and † indicate that 3D MSDS-UNet or 3D MSDS-UNet-GM significantly outperformed the comparable methods. Student's *t*-test at a level of 0.05 was used.

	2D U-Net	3D U-Net	3D U-Net with ResNet	3D MSDS-UNet	3D MSDS-UNet-GM
Dice	0.578*†	0.643*†	0.649*†	0.664†	**0.675**
Sensitivity	0.566*†	0.693*†	0.705*†	0.728†	**0.731**
*F*1-score	0.586*†	0.639*†	0.641*†	0.667†	**0.682**

Algorithm 1. Training of MSDS-UNet with GradNorm.

Initialize $\alpha_m = 0 \,\forall\, m$
Initialize network weights $\boldsymbol{W}_s$ and $\boldsymbol{W}_m \,\forall\, m$
Pick value for $\alpha > 0$
for $t = 0$**to**max_train_steps **do**
Compute $G^{(m)}_{\boldsymbol{W}_s}(t) = \|\nabla_{\boldsymbol{W}_s} \alpha_m(t) L_m(t)\|_2$ and $r_m(t) \forall m$ [L_m is the m-th loss of the corresponding the m-th layers, r_m is the relative inverse training rate of task m which can be used to rate balance our gradients]
Compute $\overline{G}_{\boldsymbol{W}_s}(t)$ by averaging the $G^{(m)}_{\boldsymbol{W}_s}(t)$
Compute $L_{\text{grad}} = \sum_m |G^{(m)}_{\boldsymbol{W}_s}(t) - \overline{G}_{\boldsymbol{W}_s}(t) \times [r_m(t)]^{\alpha}|$
Compute GradNorm gradients $\nabla_{\alpha_m} L_{\text{grad}}$, keeping targets $\overline{G}_{\boldsymbol{W}_s}(t) \times [r_m(t)]^{\alpha}$ constant
Compute standard gradients $\nabla_{\boldsymbol{W}_s} L(t)$
Update $\alpha_m(t) \rightarrow \alpha_m(t+1)$ using $\nabla_{\alpha_m} L_{\text{grad}}$
Update $\boldsymbol{W}_s(t) \rightarrow \boldsymbol{W}_s(t+1)$ using $\nabla_{\boldsymbol{W}_s} L(t)$[standard backward pass]
Update $\boldsymbol{W}_m(t) \rightarrow \boldsymbol{W}_m(t+1)$ using $\nabla_{\boldsymbol{W}_m} L(t)$[standard backward pass]
Renormalize $\alpha_m(t+1)$ so that $\sum_m \alpha_m(t+1) = 1$
end for

4. Experiment and results

4.1. Experiment on a local dataset from the Liaoning Cancer Hospital in China

4.1.1. Data and experimental setting

All the 220 cases of 3-dimensional CT image data used in the experiment were real data obtained from the Liaoning Cancer Hospital in Shenyang, China. The number of each patient CT series varies from 31 to 260. The height and width are both 512 in all data. The protocol of this retrospective study was approved by the Ethics of Committees of Liaoning Cancer Hospital. Informed consent was waived because of the respective nature of the study, and all the private information of patients was anonymized by the investigators after data collection. Images were obtained from a CT scanner of Philips (iCT256) with resolution of 512×512 and slice thickness of 5 mm. The pixel spacing is 0.79. The scanning conditions are120 kV and 20–50 mA. Both training and testing processes use the data after lung segmentation and all CT data are re-sampled to 1 mm × 1 mm × 1 mm. During the training phase, the images and the ground truth are divided into multiple cubic parts with the size of $128 \times 128 \times 128$. The batchsize is set to 1 because of GPU memory limitations and the initial learning rate is set to 0.001 which is multiplied by 0.1 for each iteration of 50 epochs for a total of 200. The optimization process uses the Adam optimizer. In each of 10 trials, a 3-fold nested cross validation procedure is employed to tune the regularization parameters (λ_1, λ_2 and λ_3). The range of each parameter varied from 10^{-1} to 10^3. The reported results were the best results of each method with the optimal parameters. The weight of each side α_m is set to $\{0.2, 0.2, 0.2, 0.4\}$ from the shallow layer to the high layer.

4.1.2. The comparison with the baseline methods

Experiments were performed on the original 2D U-Net, 3D U-Net, 3D U-Net w2ith ResNet and our proposed 3D MSDS-UNet and 3D MSDS-UNet-GM using a 5-fold cross-validation method. Table 1 shows the segmentation results of the four models with respect to dice score, sensitivity and *F*1-score. 3D MSDS-UNet indicates the method where the weight of each side is manually set to $\{0.2, 0.2, 0.2, 0.4\}$, and 3D MSDS-UNet-GM indicates the method with GradNorm as the optimization of the side weights. From Table 1, it can be seen that the tumor segmentation results of both the 3D MSDS-UNet models with the decision branch is better than the 2D U-Net and 3D U-Net with respect to dice score, sensitivity and *F*1-score. The result proves that the deep supervision strategy can indeed help the 3D U-Net model to identify lung tumors with higher segmentation performance. Moreover, it was observed that introducing the ResNet blocks enhances the performance of the traditional 3D U-Net. Furthermore, the optimization mechanism with GradNorm can further improve the performance, which indicates

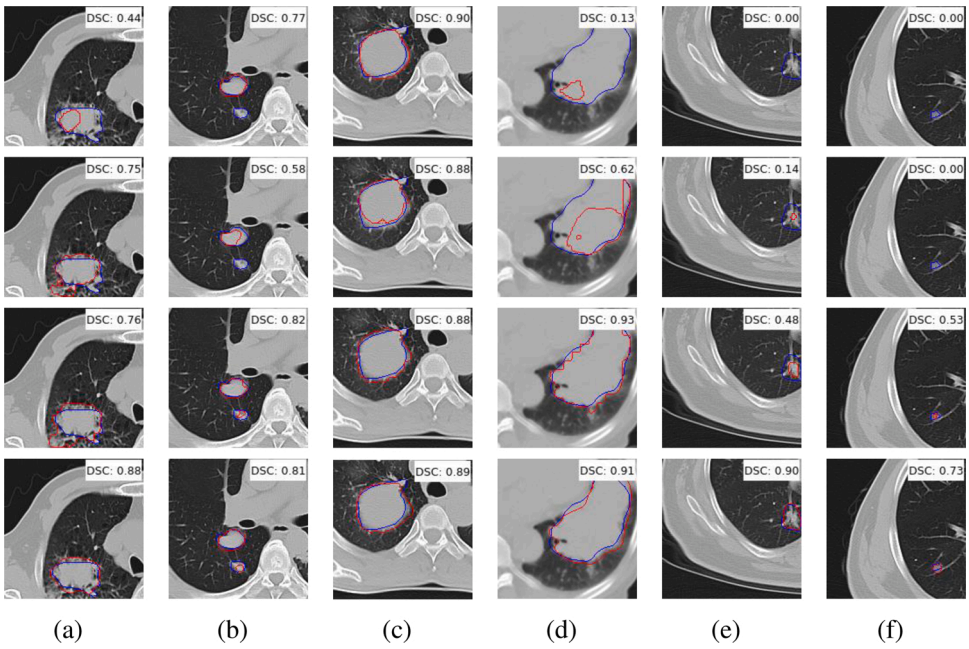


Fig. 8. Lung tumors segmentation comparison among 2D U-Net, 3D U-Net, our 3D MSDS-UNet and 3D MSDS-UNet-GM. Each column represents a patient's lung CT case and each row represents the lung tumor segmentation results of 2D U-Net, 3D U-Net, our 3D MSDS-UNet and 3D MSDS-UNet-GM respectively. The blue and red indicate the ground truth and predicted boundary, respectively.

Table 2
The comparable segmentation results of tumors with the state-of-the-art 3D segmentation methods. Superscript symbols * and † indicate that 3D MSDS-UNet or 3D MSDS-UNet-GM significantly outperformed the comparable methods. Student's *t*-test at a level of 0.05 was used.

	V-Net	3D CNN	3D MSDS-UNet	3D MSDS-UNet-GM
Dice	0.633*†	0.651*†	0.664†	**0.675**
Sensitive	0.687*†	0.724*†	0.728†	**0.731**
*F*1-score	0.648*†	0.661*†	0.667†	**0.682**

Table 3
The segmentation results of different objective functions in our MSDS-UNet. Superscript symbols * indicates that 3D MSDS-UNet significantly outperformed that method. Student's *t*-test at a level of 0.05 was used. The value in brackets indicates the *p*-value.

	Dice	Sensitivity	*F*1-score
No deep supervision	0.649* (0.066)	0.705* (0.035)	0.641* (0.031)
Only hard-fusion	0.654* (0.015)	0.711* (0.013)	0.650* (0.008)
Only soft-fusion	0.659* (0.020)	0.715* (0.028)	0.652* (0.017)
Only side supervision	0.652* (0.019)	0.694* (0.031)	0.646* (0.026)
Hard fusion + soft fusion	0.658 (0.095)	0.722(0.110)	0.659* (0.015)
Hard fusion + soft fusion + side supervision	**0.664**	**0.728**	**0.667**

that the weights of multi-task learning formulation are critical for our 3D MSDS-UNet model.

Fig. 8 shows five segmentation examples generated by the four U-Net architecture methods on some test sets for comparison. We can observe that our models with deep supervision achieve better performance. Therefore, we hypothesize that the inclusion of more supervision scales are allowing the model to distinguish the tumor boundaries better. The multi-scale context information can be learned by our network which will then facilitates the tumor segmentation in the test stage.

We empirically demonstrate state-of-the-art performance on 3D tumor segmentation in Table 2. The experimental results have indicated the effectiveness of the proposed deeply supervised 3D U-Net and suggest that our approach could acquire competitive performance as the state-of-the-art lung tumor segmentation methods. To further validate the effectiveness of our method, we compared the proposed method with different deep supervision schemes. Table 3 shows the segmentation results of different objective functions in our MSDS-UNet. From the experimental results in Table 3, we observe that all the deep supervision formulations achieve improved performance over the baseline method with no deep supervision, and the further combination of three different deep supervision mechanism performs the best among all the competing methods. With the losses calculated by the predictions from side outputs of different layers, more effective gradients back propagation can be achieved by direct supervision on the hidden layers. This experiment further demonstrates that the model with more supervision scheme achieve a better performance. Exploiting the deep supervision may be advantageous for tumor representation learning, resulting in a better segmentation performance. The proposed two-pathway multi-level deep supervision can make use of the complementary multi-scale features for final prediction from local and global perspectives, thus is more effective.

4.1.3. *The influence of deep supervision mechanism*

To further validate the effectiveness of deeply supervision in our method, we also systematically analyze the impact of different setting of supervision on the segmentation performance. We run several ablation experiments to explore the best side output settings.

Table 4
The average segmentation results of our 3D MSDS-UNet with different amounts of sides (α indicates the weight of corresponding sides). Superscript symbols * indicates that 3D MSDS-UNet significantly outperformed that method. Student's *t*-test at a level of 0.05 was used. The value in brackets indicates the *p*-value.

	(4)-side	(34)-side	(234)-side	(1234)-side
α	{1}	{0.2, 0.8}	{0.2, 0.2, 0.6}	{0.2, 0.2, 0.2, 0.4}
Dice	0.653* (0.009)	0.649* (0.024)	0.654* (0.031)	**0.664**
Sensitivity	0.701* (0.022)	0.713* (0.019)	0.718* (0.013)	**0.728**
*F*1-score	0.649* (0.026)	0.644* (0.017)	0.650* (0.008)	**0.667**

Table 5
The average segmentation results of our 3D MSDS-UNet with three different combination of sides. Superscript symbols * indicates that 3D MSDS-UNet significantly outperformed that method. Student's *t*-test at a level of 0.05 was used. The value in brackets indicates the *p*-value.

	(234)-side	(134)-side	(124)-side
Dice	0.654* (0.014)	0.657* (0.012)	**0.673**
Sensitivity	0.718* (0.022)	0.680* (0.022)	**0.743**
*F*1-score	0.650* (0.019)	0.653* (0.034)	**0.669**

Table 6
The average segmentation results of large-scale tumors by different models. Superscript symbols * indicates that 3D MSDS-UNet significantly outperformed that method. Student's *t*-test at a level of 0.05 was used.

Method	Dice	Sensitivity	*F*1-score
2D U-Net	0.673*	0.645*	0.679*
3D U-Net	0.735*	0.745*	0.730*
3D MSDS-UNet with (34)-side	0.749	0.784	0.733
3D MSDS-UNet with (234)-side	**0.768**	0.787	**0.754**
3D MSDS-UNet with (134)-side	0.738	0.734*	0.733
3D MSDS-UNet with (124)-side	0.758	**0.790**	0.753
3D MSDS-UNet with (1234)-side	0.743	0.764	0.739

(1) The influence of different amounts of deep supervision.

From the experimental results in Table 4, it can be observed that (1234)-side model with all supervision working improved performance with the comparable methods. This means that multi-scale information is complementary and more supervision for side outputs can bring in additional performance gain. However, the exception is that the (34)-side one performs worse than the 4-th side with fewer supervision in terms of the dice score and *F*1-score. It indicates that the inappropriate combination can deteriorate the segmentation performance.

(2) The influence of the deep supervision with different side outputs.

We make a comparison among the supervisions with the same side number (3) but different positions. We found that the (124)-side model achieves a best performance from the segmentation shown in Table 5. The reason is that the outputs of branch with coarser layers are more important. The coarser side outputs capture rich spatial information. They are capable of successfully highlighting the boundaries of tumors, especially for the small tumors. To our surprise, the (124)-side model performs better than the (1234)-side one which tells us that the 3rd side negatively influences our model, and the more side outputs reduce the weights of the prediction of the appropriate layer side, resulting in lower segmentation performance. Both results demonstrate that the supervision with the appropriate number and side position is critical for the segmentation performance.

4.1.4. *The performance of 3D MSDS-UNet on the tumors at different scales*

Lung tumors have various scale and we quantitatively demonstrate

Table 7
The average segmentation results of medium-scale tumors by different models. Superscript symbols * indicates that 3D MSDS-UNet significantly outperformed that method. Student's *t*-test at a level of 0.05 was used.

Method	Dice	Sensitivity	*F*1-score
2D U-Net	0.621*	0.608*	0.612*
3D U-Net	0.665*	0.709*	0.630*
3D MSDS-UNet with (34)-side	0.676*	0.718	0.672
3D MSDS-UNet with (234)-side	0.693	0.758	0.687
3D MSDS-UNet with (134)-side	**0.698**	0.724*	**0.694**
3D MSDS-UNet with (124)-side	0.695	**0.759**	0.691
3D MSDS-UNet with (1234)-side	0.682	0.742	0.678

Table 8
The average segmentation results of small-scale tumors by different models. Superscript symbols * indicates that 3D MSDS-UNet significantly outperformed that method. Student's *t*-test at a level of 0.05 was used.

Method	Dice	Sensitivity	*F*1-score
2D U-Net	0.477*	0.502*	0.477*
3D U-Net	0.519*	0.619*	0.515*
3D MSDS-UNet with (34)-side	0.522*	0.644*	0.515*
3D MSDS-UNet with (234)-side	0.499*	0.600*	0.532*
3D MSDS-UNet with (134)-side	0.526*	0.573*	0.564*
3D MSDS-UNet with (124)-side	0.556	**0.676**	0.568*
3D MSDS-UNet with (1234)-side	**0.558**	0.651	**0.585**

significant improvements on the different scale of tumors in this experiment. The results of 3D MSDS-UNet models with different amounts of side output, 2D U-Net and 3D U-Net are shown in Tables 6–8.

As shown in Tables 6–8, 3D MSDS-UNet outperforms the base U-Net architecture in segmenting all scales of tumors, especially for small tumors. The task of the segmentation on the small size tumor is the most difficult. The multi-scale context information is needed to be learned by our network which then facilitates the target segmentation in the test stage. For the large and medium scale tumors, the 3D MSDS-UNet with three sides is best. The 3D MSDS-UNet with four sides is even worse that the one with two sides in terms of dice score and sensitivity. The identification of large tumors depends on the advanced features of the top feature maps. The supervision of the shallow hidden layers is not necessary. The more sides output reduce the weights of the prediction of the higher layer side, resulting in lower segmentation performance. For the small scale one, the 3D MSDS-UNet with (1234)-side outperforms the other versions. Besides it, the 3D MSDS-UNet with (123)-side is the second best algorithm. The reason is that the recognition of small tumors depends on the underlying features contained in the shallow hidden layers. Therefore, achieving an effective feature representation with direct supervision in the lower layer is important. Moreover, the importance of the segmentation output of branches are different for the tumors with variable scales. However, in our current work, the weights for branches in the deep supervision are fixed. In the future, we will investigate an adaptive weighting scheme for different tumors.

In order to investigated the contribution of the multi-level multi-scale supervision. We chose two small scale tumors as examples. From Figs. 9 and 10, we can find that the inclusion of more supervision scales allows the U-Net to distinguish the boundaries better. We observe that the dice score is zero on some slice images for only 4-side or only (34)-side. The result demonstrates that the deep supervision in our 3D MSDS-UNet can captures the features of different scale, thus sensitive in small

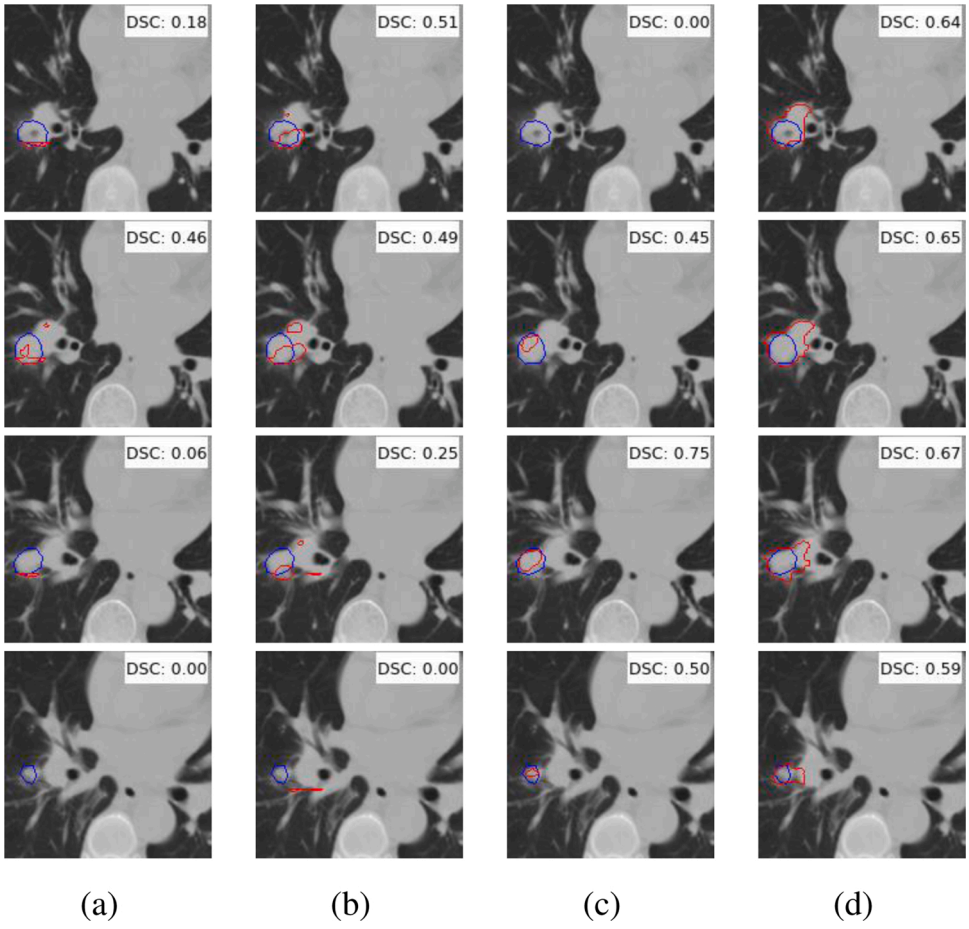


Fig. 9. The results of example 1. The each row indicates the results of different slices from the same case by different algorithms. (a) 4-side; (b) the combination of (34)-side; (c) the combination of (234)-side; (d) the combination of (1234)-side. The blue and red indicate the ground truth and predicted boundary, respectively.

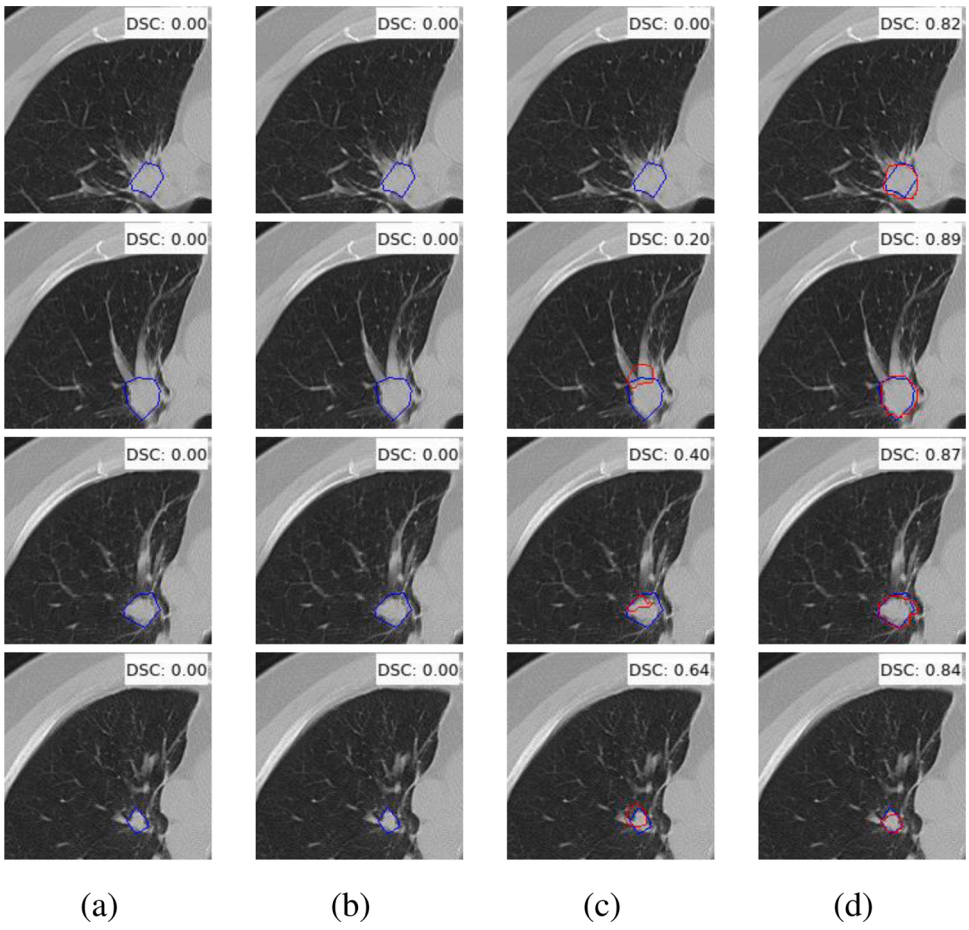


Fig. 10. The results of example 2.The each row indicates the results of different slices from the same case by different algorithms. (a) 4-side; (b) the combination of (34)-side; (c) the combination of (234)-side; (d) the combination of (1234)-side. The blue and red indicate the ground truth and predicted boundary, respectively.

tumors. Small tumors may have lost their response when the feature map has reached a certain depth, which undoubtedly makes it more difficult for these methods to detect small tumors accurately. Only the features from the top layers fail to achieve expected segmentation performance due to the loss of spatial details in the bottom sides. The direct deep supervision on the earlier layers can improve the segmentation of small objects. The deep subversion can promote the fusion of semantic meaningful information in the top sides and the complementary spatial details in the bottom sides, leading to break through the bottleneck of U-Net based deep learning segmentation model for multiple scale tumors, especially small tumors. This experiment further demonstrates that exploiting the multi-level multi-scale supervision resulted in better segmentation performance.

4.2. Experiment on the public data from TCIA

4.2.1. Data and experimental setting

This collection contains images from 422 non-small cell lung cancer (NSCLC) patients. For these patients pretreatment CT scans, manual delineation by a radiation oncologist of the 3D volume of the gross tumor volume and clinical outcome data are available. We used the TCIA dataset (The Cancer Imaging Archive, http://cancerimagingarchive.net/) as the training cohort as it had the largest tumors and contained many difficult to detect tumors such as those attached to the mediastinum and the chest wall. This dataset, called Lung1, contains data of patients treated at the MAASTRO Clinic, Netherlands, previously identified by the Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) and made publicly available for download. Table 9 details the characteristics of the datasets and the patient demographics for three-folds cross validation.

Table 9
Number of patients and training patches for training, validation and test in the dataset.

Fold ID	Data	Num of patients	Tumor	Without tumor
Fold 1	Train	296	5126	30730
	Validation	60	1091	6414
	Test	66	1015	6129
Fold 2	Train	297	5260	30972
	Validation	60	1071	6012
	Test	65	901	6289
Fold 3	Train	296	5167	30343
	Validation	60	1034	6721
	Test	66	1031	6209

4.2.2. The comparison with state-of-the-art methods

We compare our proposed method with several state-of-the-art methods available in the literature for performing segmentation tasks.

FRRN (Pohlen et al., 2017): It is a ResNet-like architecture unites strong recognition performance with precise localization capabilities by combining two distinct processing streams. One stream undergoes a sequence of pooling operations and is responsible for understanding large-scale relationships of image elements; the other stream carries feature maps at the full image resolution, resulting in precise boundary

Table 10
The segmentation results of tumors with the different segmentation models on the only slices with containing lung tumors. Superscript symbols * and † indicates that 3D MSDS-UNet or 3D MSDS-UNet-GM significantly outperformed the comparable methods. Student's *t*-test at a level of 0.05 was used.

	3D U-Net with ResNet	RD-UNet	FRRN	In-MRRN	MSDS-UNet	MSDS-UNet-GM
Dice	0.667*†	0.588*†	0.516*†	0.564*†	0.683†	**0.691**
Sensitivity	0.728*†	0.60*†	0.703	0.665*†	0.738†	**0.746**
Precision	0.706*†	0.691†	0.489*†	0.564*†	0.717†	**0.719**
*F*1-score	0.688*†	0.606*†	0.537*†	0.582*†	0.709†	**0.724**
False Positives	894	888	5164	4860	790	**765**
False Negatives	134	231	**58**	93	175	141

Table 11
The segmentation results of tumors with different segmentation models on the all slices. Superscript symbols * and † indicates that 3D MSDS-UNet or 3D MSDS-UNet-GM significantly outperformed the comparable methods. Student's *t*-test at a level of 0.05 was used.

	3D U-Net with ResNet	RD-UNet	FRRN	In-MRRN	MSDS-UNet	MSDS-UNet-GM
Dice	0.516*†	0.397*†	0.168*†	0.167*†	0.538	**0.552**
Sensitivity	0.714*†	0.562*†	**0.726**	0.658*†	0.707	0.709
Precision	0.472*†	0.368*	0.117*†	0.115*†	0.498	**0.503**
*F*1-score	0.513	0.391*†	0.166*†	0.167*†	0.521	**0.529**

adherence.

Incremental MRRN (In-MRRN) (Jiang et al., 2018a) is an extension of FRRN by residually combining features computed at multiple image resolutions, whereby a dense feature representation is computed by simultaneously combining feature maps at multiple image resolutions and feature levels. Such a dense feature representation increases the network capacity and ultimately enables the network to recover the input image spatial resolution better than the existing methods. The method was proposed for volumetrically segmenting lung tumors.

Recurrent 3D-DenseUNet (RD-UNet) (Kamal et al., 2018): The proposed architecture consists of a 3D encoder block that learns to extract ne-grained spatial and coarse-grained temporal features, a recurrent block of multiple Convolutional Long Short-Term Memory (ConvLSTM) layers to extract ne-grained spatio-temporal information, and finally a 3D decoder block to reconstruct the desired volume segmentation masks from the latent feature space. The author's team obtained first Runner-Up in the 2018 VIP Cup challenge of which the dataset is from TCIA dataset.

Quantitative results from the methods of FRRN, incremental MRRN, recurrent 3D-DenseUNet and MSDS-UNet on the TCIA data are presented in Tables 10 and 11. From Tables 10 and 11, we can see that the proposed methods consistently achieves better segmentation performance than the competing methods in terms of dice score, sensitivity and *F*1-score on both the only slices with containing tumors and all slices individually, which demonstrates the effectiveness of our MSDS-UNet method. With the student's *t*-test at a level of 0.05, our proposed method significantly outperforms the contenders on the most cases. Moreover, our method obtains the least false positive instances. Compared with Incremental MRRN, our MSDS-UNet performs the tumor segmentation with considering the 3D tumor structure and deep supervision mechanism to comprehensively address these challenges of volumetric medical image segmentation. Moreover, it can be observed that the performances of Incremental MRRN is poor when tested on the all slices, since the amount of false positive is large without taking full advantage of the special information encoded in the volumetric data. The high variations impose more difficulty in extracting potential features. Only fusing the feature maps from multiple image resolutions is not beneficial. There are still potential issues in training a deep network with inadequate discriminative power towards learned features and exploding or vanishing gradients. Especially when the depth of the model increases, the multi-level features cannot be learned well before fusing due to lacking of the appropriate supervision. From the results, we can find that with two-pathway deeply supervision from both global and local perspectives, more inherent features from different scales can be achieved. Finally, we empirically validate the significance of our thresholding and morphological operation. From Table 12, we can see how the performance deteriorates without any thresholding or morphological operations.

Table 12
Performance of our proposed method for different thresholds.

Model	Mean Dice	Median Dice	False Positives	False Negatives
0.5 threshold	0.688	0.767	756	169
0.6 threshold	0.694	0.769	661	191
0.7 threshold	0.682	0.755	545	222
0.8 threshold	0.659	0.759	489	243
0.9 threshold	0.631	0.736	429	281
No threshold	0.621	0.659	4185	3

Examples of segmentation obtained with the comparable algorithms are shown in Fig. 11. MSDS-UNet behaves very well in preserving the hierarchical structure of the tumor, which highlighted that combining the 3D module with deep supervision in the U-Net architecture is a promising approach for semantic medical image segmentation. It is capable of precise segmentation for different scales and locations of lesions. The columns show the segmentation results in the following order: 3D U-Net with ResNet, Incremental MRRN, Recurrent 3D-DenseUNet and 3D MSDS-UNet. Fig. 12 illustrates the convergence of 3D U-Net with ResNet and MSDS-UNet. When comparing the learning curves of the 3D MSDS-UNet and the 3D U-Net with ResNet, the 3D MSDS-UNet converges much faster than the original 3D U-Net. These results demonstrate the proposed 3D deep supervision mechanism can effectively speed up the training procedure by overcoming optimization difficulties through managing the training of the lower layers in the network.

From Table 12, it can be found that the deep supervision mechanism can effectively cope with the optimization problem of gradients vanishing or exploding when training a 3D deep model, accelerating the convergence speed. In addition, MSDS-UNet is measured for its computational requirements in layer number, parameter number and inference speed and compared with the comparable models (Table 13). From the results in Table 13, MSDS-UNet requires slightly more number of parameters compared with 3D U-Net with ResNet, FRRN and Incremental MRRN since our model is a 3D model while both RRN and Incremental MRRN are 2D models. The computational complexity is an inherent disadvantage of 3D network model. However, a large margin improvement has been achieved in terms of segmentation performance in Tables 10 and 11. Compared with RD-UNet which is also a 3D model, MSDS-UNet has a lower computational complexity. Therefore our model achieves a good trade-off between effectiveness and efficiency.

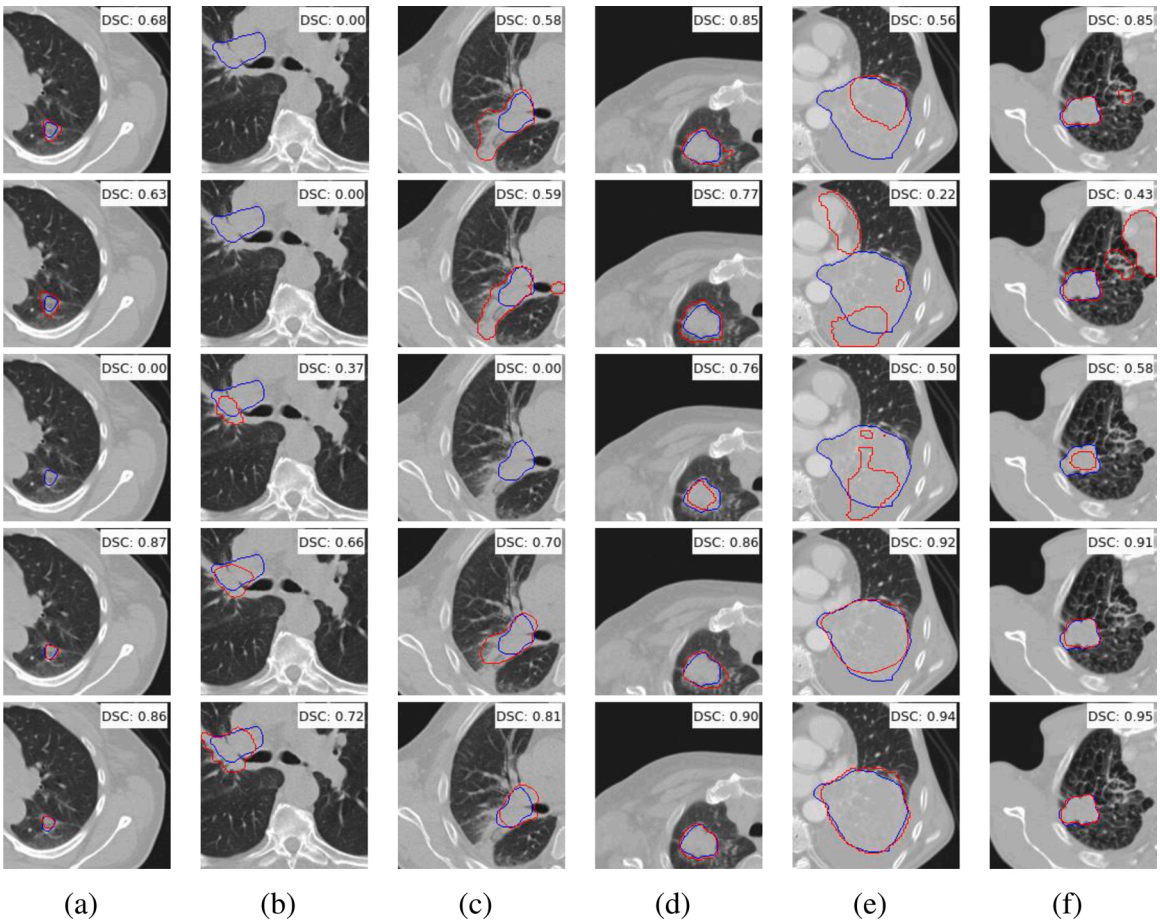


Fig. 11. Lung tumors segmentation comparison among 3D U-Net with ResNet, incremental MRRN, recurrent 3D-DenseUNet, our 3D MSDS-UNet and 3D MSDS-UNet-GM. Each column represents a patient's lung CT case and each row represents the lung tumor segmentation result of 3D U-Net with ResNet, incremental MRRN, recurrent 3D-DenseUNet, our 3D MSDS-UNet and 3D MSDS-UNet-GM, respectively. The blue and red indicate the ground truth and predicted boundary, respectively.

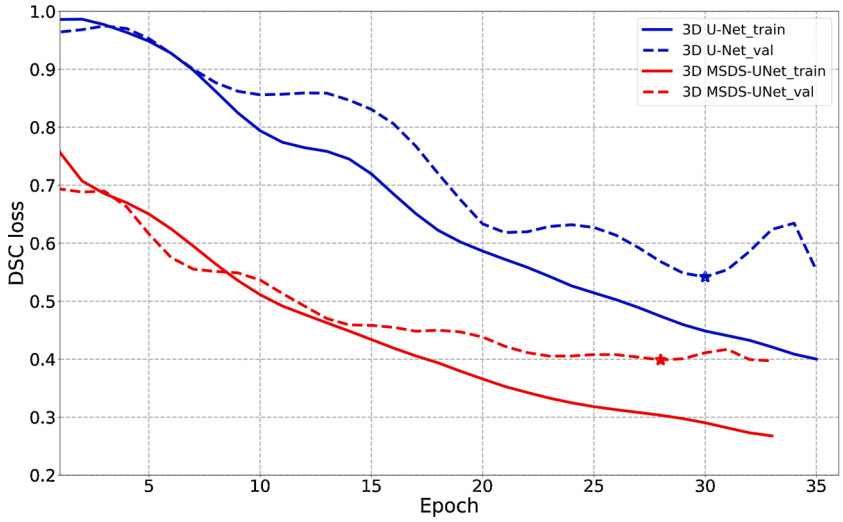


Fig. 12. The learning curves of two methods.

4.3. Experiment on a local dataset of COVID-19 patients from China

Accurate and rapid diagnosis of COVID-19 suspected cases plays a crucial role in timely quarantine and medical treatment, which is also of great importance for patients' prognosis. In this work, we evaluated our MSDS-UNet for automatic segmentation of pathologic COVID-19 associated tissue areas from clinical CT images available from a dataset with 108 cases in China.

In our experiments, 108 cases of 3-dimensional lung CT image data from different COVID-19 patients that treated in different hospitals were used for model training and testing. 58 patients from Harbin (Heilongjiang Province) were scanned using a 256-slice CT scanner (Philips

Table 13
The network parameters of MSDS-UNet and the comparable methods.

	3D U-Net with ResNet	RD-UNet	FRRN	Incremental MRRN	MSDS-UNet
Layer number	59	20	49	56	62
Parameters number	26.38 M	29.10 M	24.83 M	28.66 M	26.45 M
Inference time	16.1 s	18.93 s	11.7 s	13.5 s	16.3 s

Table 14
The segmentation results of COVID-19 lung infection with different U-Net based segmentation models. Superscript symbols * and † indicates that 3D MSDS-UNet or 3D MSDS-UNet-GM significantly outperformed the comparable methods. Student's *t*-test at a level of 0.05 was used.

	U-Net	U-Net with ResNet	COVID-CT-Mask-Net	Inf-Net	MSDS-UNet	MSDS-UNet-GM
Dice	0.630*†	0.667*†	0.628*†	0.658†	0.671	**0.674**
Sensitivity	0.727*†	0.733*†	0.644*†	0.704†	0.710†	**0.733**
Precision	0.646*†	0.700*†	0.688*†	0.713†	0.714†	**0.727**

Healthcare, Cleveland, OH, US), 24 patients from Shuangyashan (Heilongjiang Province)were scanned with Somatom Balance CT (Siemens Healthcare, Forchheim, Germany), 16 patients from Uygur autonomous region (Xinjiang Province) were examined with LightSpeed Plus (GE, Medical System, Milwaukee, USA), 10 patients from Chengdu (Sichuan Province) were examined with 128-slice DSCT (Siemens Healthcare, Forchheim, Germany). All these CT images were reconstructed into a slice thickness of 1.0–2.0 mm. Scan were performed in the supine position during end-inspiration. All CT images were de-identified before sending for analysis. This study is in compliance with the Institutional Review Board of each participating institutes. Informed consent was exempted by the IRB because of the retrospective nature of this study.

Anam-Net (Paluru et al., 2021): is a network architecture utilized for segmenting abnormalities in COVID-19 chest CT images. Fully convolutional anamorphic depth blocks (AD-blocks) with depthwise squeezing and stretching have been incorporated after downsampling operation in the encoder and decoder.

Inf-Net (Fan et al., 2020a) is a COVID-19 lung CT infection segmentation network, which utilizes an implicit reverse attention and explicit edge-attention to improve the identification of infected regions.

Table 14 and Fig. 13 report our results and the comparisons with the other state-of-art networks quantitatively and qualitatively. Both MSDS-UNet methods achieve consistent improvement in terms of Dice, sensitivity and precision. This is because lesions appear at varying scales in CT slices; and thus, a multi-scale approach using all segmentation branches with deep supervision is essential for accurate segmentation.

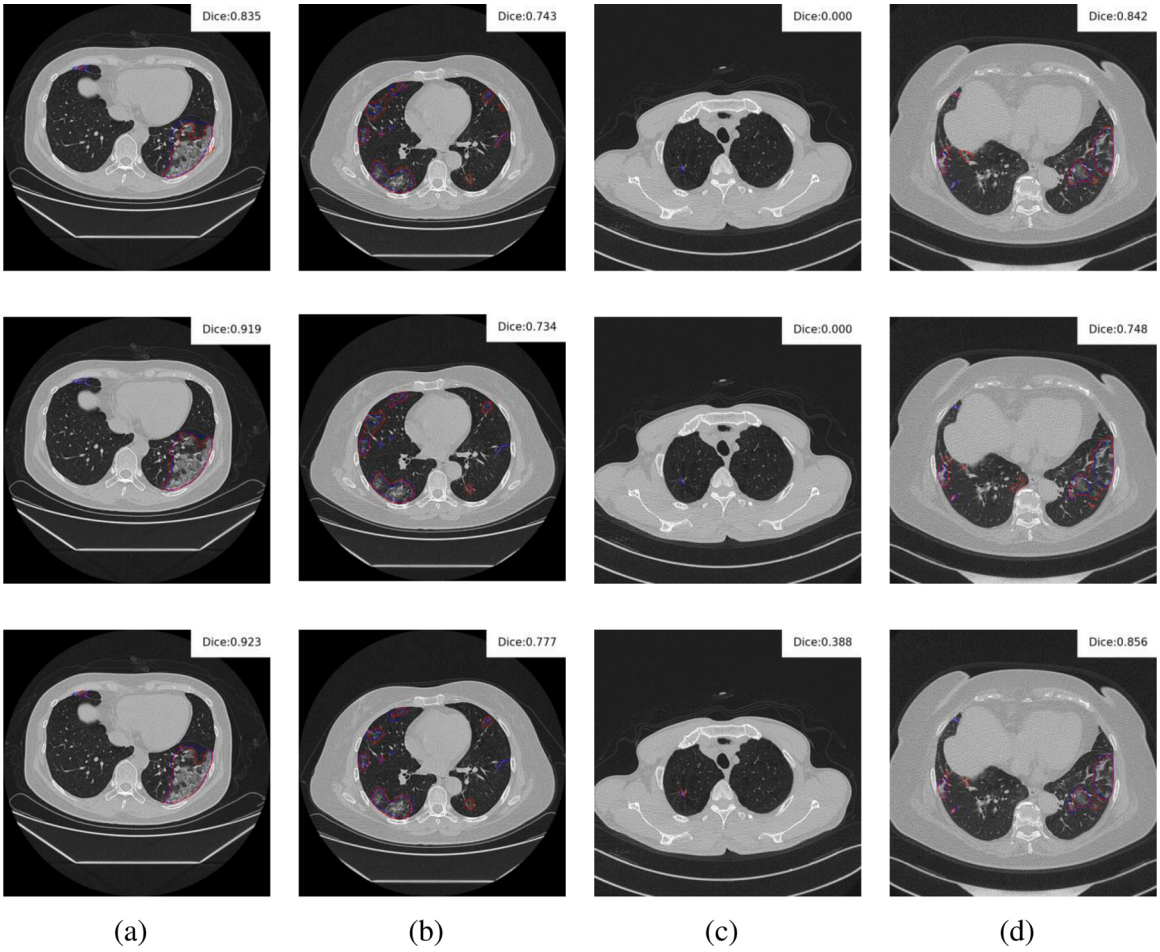


Fig. 13. Lung tumors segmentation comparison among COVID-CT-Mask-Net, Inf-Net and our 3D MSDS-UNet. Each column represents a patient's lung CT case and each row represents the lung tumor segmentation results of COVID-CT-Mask-Net, Inf-Net and 3D MSDS-UNet respectively. The blue and red indicate the ground truth and predicted boundary, respectively.

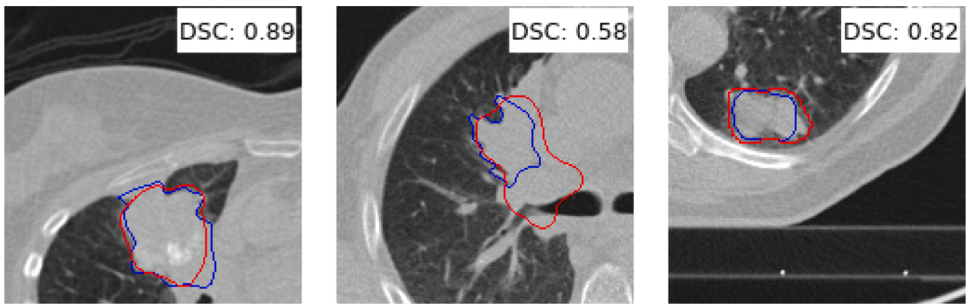


Fig. 14. Some inaccurately segmented cases.

5. Limitation and the future research

To further strengthen MSDS-UNet, the current work presents several extensions to our previous work:

1. Through some inaccurately segmented cases (Fig. 14), we can observe that our current model ignores the learning of the contour regions. To derive the essential contour, which acts as a complementary feature with the global shape feature. During lung tumor or infection detection, clinicians first roughly locate an infected region and then accurately extract its contour according to the local appearances. We therefore argue that the area and boundary are two key characteristics that distinguish normal tissue and abnormal one. In the next research, we guide our segmentation network model to learn complementary contour region that can aid the accurate delineation of the target object.

2. We will investigate the extensibility of MSDS-UNet to multiple advanced encoder backbones, such as HRNet (Sun et al., 2019) or UNet++ (Zhou et al., 2019).

6. Conclusion

The volumetric segmentation of lung tumor is a challenging problem since multi-scale of target tumors limit the traditional U-Net. In this paper, we presented a 3D deeply supervised network for lung tumor segmentation. We propose a 3D deep supervision mechanism by formulating an objective function that directly guides the training of the hidden layers in order to reinforce the propagation of gradients flows within the network and hence learn more powerful and representative features. Different from previous deeply supervision architectures, a combination of the side-output fused predictions for the final prediction with the ensemble of multi-scale predictions of multi-resolution segmentation map, to increase the network's capacity for learning richer representations of lung tumors from global and local perspectives. We evaluated MSDS-UNet using three medical imaging datasets covering lung tumor segmentation and COVID-19 infection segmentation. Our experiments demonstrate that the mechanism results in more accurate segmentation for multi-scale lung tumors, especially on the small scale cancer, which further verifies the merit of the proposed deep supervision mechanism.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (No. 62041601 and No. 62076059) , the Fundamental Research Funds for the Central Universities (No. N2016001) and the Beijing Municipal Science and Technology Planning Project (Grant No. Z211100003521009).

References

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. International Conference on Medical Image Computing and Computer-Assisted Intervention 424–432.

Albarazanchi, H.A., Qassim, H., Verma, A., 2016. Novel CNN Architecture with Residual Learning and Deep Supervision for Large-Scale Scene Image Categorization doi: 10.1109/UEMCON.2016.7777858.

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.

Chen, C., Qi, F., 2018. Single image super-resolution using deep CNN with dense skip connections and inception-ResNet. 2018 9th International Conference on Information Technology in Medicine and Education (ITME).

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, Z., Badrinarayanan, V., Lee, C.-Y., Rabinovich, A., 2018. GradNorm: gradient normalization for adaptive loss balancing in deep multitask networks. International Conference on Machine Learning 794–803.

Christ, P.F., Elshaer, M.E.A., Ettlinger, F., Tatavarty, S., Bickel, M., Bilic, P., Rempfler, M., Armbruster, M., Hofmann, F., D'Anastasi, M., 2016. Automatic liver and lesion segmentation in CT using cascaded fully convolutional neural networks and 3d conditional random fields. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham. doi:10.1007/978-3-319-46723-8_48.

Chung, M., Lee, J., Lee, M., Lee, J., Shin, Y.-G., 2020. Deeply self-supervised contour embedded neural network applied to liver segmentation. Comput. Methods Progr. Biomed. 192, 105447.

Dou, Q., Yu, L., Chen, H., Jin, Y., Yang, X., Qin, J., Heng, P.-A., 2017a. 3D deeply supervised network for automated segmentation of volumetric medical images. Med. Image Anal. 41, 40–54.

Dou, Q., Yu, L., Chen, H., Jin, Y., Yang, X., Qin, J., Heng, P.-A., 2017b. 3D deeply supervised network for automated segmentation of volumetric medical images. Med. Image Anal. 41, 40–54.

Fan, D.-P., Zhou, T., Ji, G.-P., Zhou, Y., Chen, G., Fu, H., Shen, J., Shao, L., 2020a. Inf-Net: automatic COVID-19 lung infection segmentation from CT images. IEEE Trans. Med. Imaging 39 (8), 2626–2637.

Fan, T., Wang, G., Li, Y., Wang, H., 2020b. MA-Net: a multi-scale attention network for liver and tumor segmentation. IEEE Access 8, 179656–179665.

Guofeng, T., Yong, L., Huairong, C., Qingchun, Z., Huiying, J., 2018. Improved U-Net network for pulmonary nodules segmentation. Optik. https://doi.org/10.1016/j.ijleo.2018.08.086.

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.

Hoffman, P.C., Mauer, A.M., Vokes, E.E., 2000. Lung cancer. Lancet 355 (9202), 479–485. https://doi.org/10.1016/S0140-6736(00)82038-3. Erratum in: Lancet 2000 Apr 8;355(9211): 1280. PMID: 10841143.

Hossain, S., Najeeb, S., Shahriyar, A., Abdullah, Z.R., Haque, M.A., 2019. A pipeline for lung tumor detection and segmentation from CT scans using dilated convolutional neural networks. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 1348–1352.

Huang, Q., Sun, J., Ding, H., Wang, X., Wang, G., 2018. Robust liver vessel extraction using 3D U-Net with variant dice loss function. Comput. Biol. Med. 101, 153–162.

Ibtehaz, N., Rahman, M.S., 2019. Multiresunet: rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural Netw. https://doi.org/10.1016/j.neunet.2019.08.025.

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H., 2017. Brain tumor segmentation and radiomics survival prediction: contribution to the brats 2017 challenge. International MICCAI Brainlesion Workshop 287–297.

Jiang, J., Hu, Y.-C., Liu, C.-J., Halpenny, D., Hellmann, M.D., Deasy, J.O., Mageras, G., Veeraraghavan, H., 2018a. Multiple resolution residually connected feature streams for automatic lung tumor segmentation from CT images. IEEE Trans. Med. Imaging 38 (1), 134–144.

Jiang, J., Hu, Y.-C., Tyagi, N., Zhang, P., Rimner, A., Mageras, G.S., Deasy, J.O., Veeraraghavan, H., 2018b. Tumor-aware, adversarial domain adaptation from CT to

MRI for lung cancer segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention 777–785.
Kamal, U., Rafi, A.M., Hoque, R., Hasan, M., et al., 2018. Lung Cancer Tumor Region Segmentation Using Recurrent 3D-DenseUNet arXiv preprint arXiv:1812.01951.
Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. 36, 61–78.
Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z., 2015. Deeply-supervised nets. Artificial Intelligence and Statistics 562–570.
Li, X., Hao, C., Xiaojuan, Q., Qi, D., Chi-Wing, F., Pheng-Ann, H., 2018. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. IEEE Trans. Med. Imaging, 1-1, doi:10.1109/TMI.2018.2845918.
Li, H., Li, A., Wang, M., 2019. A novel end-to-end brain tumor segmentation method using improved fully convolutional networks. Comput. Biol. Med. 108, 150–160.
Liu, S., Liang, Y., Gitter, A., 2019. Loss-balanced task weighting to reduce negative transfer in multi-task learning. Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33 9977–9978.
Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 3431–3440.
Ma, C., Luo, G., Wang, K., 2018. Concatenated and connected random forests with multiscale patch driven active contour model for automated brain tumor segmentation of MR images. IEEE Trans. Med. Imaging 37 (8), 1943–1954.
Milletari, F., Navab, N., Ahmadi, S.-A., 2016. V-Net: fully convolutional neural networks for volumetric medical image segmentation. 2016 Fourth International Conference on 3D Vision (3DV), IEEE 565–571.
Norman, B., Pedoia, V., Majumdar, S., 2018. Use of 2D U-Net convolutional neural networks for automated cartilage and meniscus segmentation of knee MR imaging data to determine relaxometry and morphometry. Radiology 288 (1), 177–185.
Paluru, N., Dayal, A., Jenssen, H.B., Sakinis, T., Cenkeramaddi, L.R., Prakash, J., Yalavarthy, P.K., 2021. Anam-Net: Anamorphic depth embedding-based lightweight CNN for segmentation of anomalies in COVID-19 chest CT images. IEEE Trans. Neural Netw. Learn. Syst.
Pohlen, T., Hermans, A., Mathias, M., Leibe, B., 2017. Full-resolution residual networks for semantic segmentation in street scenes. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 4151–4160.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention 234–241.
Sun, K., Xiao, B., Liu, D., Wang, J., 2019. Deep high-resolution representation learning for human pose estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 5693–5703.
Tureckova, A., Turecek, T., Kominkova Oplatkova, Z., Rodriguez-Sanchez, A.J., 2020. Improving CT image tumor segmentation through deep supervision and attentional gates. Front. Robot. AI 7, 106.
Xu, Y., Gong, M., Fu, H., Tao, D., Zhang, K., Batmanghelich, K., 2018. Multi-scale masked 3-D U-Net for brain tumor segmentation. International MICCAI Brainlesion Workshop 222–233.
Yang, B., Xiang, D., Yu, F., Chen, X., 2018. Lung tumor segmentation based on the multi-scale template matching and region growing. In: Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging, vol. 10578. International Society for Optics and Photonics, p. 105782Q.
Zeiler, M.D., Krishnan, D., Taylor, G.W., Fergus, R., 2010. Deconvolutional networks. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. doi:10.1109/CVPR.2010.5539957.
Zeng, G., Yang, X., Li, J., Yu, L., Heng, P.-A., Zheng, G., 2017. 3D U-Net with multi-level deep supervision: fully automatic segmentation of proximal femur in 3D MR images. International Workshop on Machine Learning in Medical Imaging 274–282.
Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2019. Unet++: redesigning skip connections to exploit multiscale features in image segmentation. IEEE Trans. Med. Imaging 39 (6), 1856–1867.
Zhu, Q., Du, B., Turkbey, B., Choyke, P.L., Yan, P., 2017. Deeply-supervised CNN for prostate segmentation. 2017 International Joint Conference on Neural Networks (IJCNN), IEEE 178–184.