Accurate Lung Nodule Segmentation With Detailed Representation Transfer and Soft Mask Supervision

Changwei Wang, Rongtao Xu, Shibiao Xu, Member, IEEE, Weiliang Meng, Member, IEEE, Jun Xiao, Member, IEEE, and Xiaopeng Zhang, Member, IEEE

Abstract—Accurate lung lesion segmentation from computed tomography (CT) images is crucial to the analysis and diagnosis of lung diseases, such as COVID-19 and lung cancer. However, the smallness and variety of lung nodules and the lack of high-quality labeling make the accurate lung nodule segmentation difficult. To address these issues, we first introduce a novel segmentation mask named “soft mask,” which has richer and more accurate edge details description and better visualization, and develop a universal automatic soft mask annotation pipeline to deal with different datasets correspondingly. Then, a novel network with detailed representation transfer and soft mask supervision (DSNet) is proposed to process the low-resolution images of lung nodules into high-quality segmentation results. Our DSNet contains a special detailed representation transfer module (DRTM) for reconstructing the detailed representation to alleviate the small size of lung nodules images and an adversarial training framework with soft mask for further improving the accuracy of segmentation. Extensive experiments validate that our DSNet outperforms other state-of-the-art methods for accurate lung nodule segmentation, and has strong generalization ability in other accurate medical segmentation tasks with competitive results. Besides, we provide a new challenging lung nodules segmentation dataset for further studies (https://drive.google.com/file/d/1SNKvDTb_0Ku0IoPsNMMHczJRI TH1Oi1wm/view?usp=sharing).

Index Terms—Detailed representation transfer, lung nodules segmentation, medical images segmentation, soft mask.

I. INTRODUCTION

SINCE December 2019, the world has suffered a severe health crisis: COVID-19 pandemic [1]. COVID-19 usually appears as a ground glass opacity (GGO) lung nodule on computed tomography (CT) images [2]. These GGO and other types of lung nodules potentially contain the risk of lung cancer, which is the deadliest type of cancer worldwide for human with a relatively low five-year survival rate of 18% [3]. Accurate segmentation of the lung nodule is of great significance for semiautomated disease screening, diagnosis, analysis, and treatment evaluation [4], [5]. However, accurate lung nodule segmentation is a challenging problem due to the following various reasons.

“One cannot make bricks without straw” is a commonly used expression to imply that it will be challenging to complete the work without the support of the primary conditions. Unfortunately, this tends to be the case for the current accurate segmentation of lung nodules without high-quality mask annotations. High-quality mask annotation datasets for accurate segmentation of lung nodules are absent, because manually labeling lung nodules is a time-consuming work obviously [6], while lung nodules annotation by radiologists is a highly subjective task, often influenced by individual bias and clinical experiences. The binary masks are obtained by relying on a pixel-by-pixel manual annotation. As shown in Fig. 1, binary masks are coarse and lacking in detail compared with soft masks. On the one hand, as shown in Fig. 1(a), the wide variety of nodule shapes leads to the challenge of accurately labeling nodules using binary masks. It can be seen that the complex edges are represented by smooth borders in the binary masks, thus losing detail information. On the other hand, as shown in Fig. 1(b), there are apparent differences in the labels of four doctors, and these inconsistencies may cause ambiguities in the segmentation network training. Cai et al. [7] proposed an automatic mask generation method using the long and short axis of response evaluation criteria in solid tumors (RECIST) [8] as weakly prior information to avoid manually labeling the mask, while this is an iterative process, and can only obtain a rough mask by relying on coarse RECIST marks and GrabCut [9], making it difficult to cope with the task of accurate segmentation of lung nodules with complex edges.

In addition, some inherent characteristics of lung nodules also cause difficulty in their segmentation. On the one hand, lung nodules are usually small (5–10 mm) [4], which results in low resolutions in images. Obviously, low-resolution loses detailed information of the lesion and makes it difficult to be segmented accurately. On the other hand, as shown in Fig. 1(a), lung nodules have significant heterogeneity. Some lung nodules, such as GGO and burr nodules, have irregular
shapes and complex edges, which also pose challenges for accurate segmentation [10]. However, as shown in Fig. 1, the existing binary mask cannot describe the edge details of these nodules well, so it is difficult to achieve accurate segmentation and visualization in this case.

To address the above issues, we propose an innovative paradigm, including an automatic accurate annotation pipeline and a segmentation network named DSNet with detailed representation transfer and soft mask supervision for accurate lung nodules segmentation. Superior to most deep learning-based lung nodules segmentation methods [11], [12], [13], [14], we not only design the segmentation network architecture but also improve the qualities of both the less-explored input (low-resolution images) and ground-truth masks (rough binary masks). Specifically, our soft mask can preserve rich edge details and smooth transition between lung nodules and surrounding pathological environments, making it have a better visualization effect and richer information. Also, we develop an automatic accurate annotation pipeline to derive accurate soft masks from different datasets. In addition, inspired by the recently popular self-supervised pretraining model [15], we carefully design a detailed representation transfer module (DRTM) in DSNet to transfer detailed representation knowledge from an off-the-shelf super-resolution generative adversarial network (SRGAN) [16] model. It can alleviate the problem of the low resolution of the input images. In general, our technical contributions have the following four aspects.

1) We introduce a new fine mask form “soft mask” and the automatic accurate labeling pipeline, which can boost segmentation accuracy and obtain excellent visualization in contrast to the traditional binary mask.

2) We design a novel DSNet with a DRTM and a soft mask-based adversarial training framework to convert low-quality images input into high-quality segmentation results. Unlike most related methods, we improve the quality of the input image and supervision mask that were rarely explored before.

3) We propose a special accurate lung nodules segmentation paradigm based on the above innovation, and its performance outperforms other state-of-the-art methods on lung nodules segmentation. In addition, it is also suitable for other accurate medical segmentation tasks after our validation.

This work extends our preliminary conference version [17] in the following aspects. First, we provide a general effective medical super-resolution (SR) model based on our self-supervised training and apply its detailed representation to promote medical segmentation task (Section III-B3a). Second, we design a novel DRTM to flexibly transfer the detailed information of the SR model into the segmentation pipeline (Section III-B3b). Third, we extend our method on new dataset to validate its robustness and generalization. Specifically, we validate the robustness by comparing models trained on cross-domain datasets with human doctors (Section IV-C) and then apply to medical segmentation datasets other than lung nodule segmentation to verify the generalization of our method (Section IV-D). Finally, we give a more inclusive and insightful discussion on our method (see Section V).

II. RELATED WORK

A. Lung Nodules Segmentation

Many classic handcrafted features-based methods have been proposed to deal with lung nodules segmentation, such as morphological operations-based methods [18], region-growing methods [19], energy optimization-based method [20], [21], and conditional random field (CRF)-based method [22]. However, these methods cannot cope with lung nodules segmentation well [14], especially for irregular-shaped nodules.

In recent years, deep learning-based methods have gained new attention for medical image analysis and processing. Some convolutional neural network (CNN)-based lung nodules segmentation methods [11], [12], [13], [14] have achieved highly competitive segmentation accuracy. In addition, some other medical segmentation methods [2], [23], [24], [25], [26], [27] have brought breakthroughs in a variety of medical segmentation tasks. More recently, some transformer-based models [28], [29] are proposed for medical image segmentation. These methods employ more complex network architecture to improve the segmentation accuracy, such as the use of attention mechanisms, spatial context, and dense connections. In contrast, our DSNet not only designs the network structure but also creatively ameliorates the quality and detail of both input images and supervised ground truth to achieve accurate lung nodules segmentation and impressive visualization.

B. Soft Labels

Soft labels have recently been applied to brain lesions segmentation [30], [31], as they are considered to have a better generalization, faster learning speed, and mitigation of network overconfidence. Specifically, Kats et al. [30] employ morphological dilation to expand the binary mask and assign

| Juxtapleural nodule | Calcific nodule | Isolated nodule | Cavitary nodule | GGO nodule | Tiny nodule |
|---------------------|----------------|----------------|----------------|------------|-------------|

Fig. 1. (a) Various types of lung nodules and manual labelings in LIDC. (b) Comparison of the labeling results of four doctors in LIDC with prediction of our proposed DSNet. Our proposed DSNet even achieves better segmentation and visualization results than human radiologists.
a fixed probability to all pixels within the expanded region to generate soft labels. Gros et al. [31] obtain soft labels by bilinear interpolation, while they still miss many edge details (e.g., small burrs around lung nodules). Both the soft labels in [30] and [31] soften the binary masks, but they are too rough to obtain accurate lung nodule segmentation with good visual effects. In contrast, the pixels of our soft mask are not discrete but continuous, meaning that our soft mask has richer and more accurate edge detail expression, which can reduce the impact of imprecise boundary annotation. Besides, our soft mask can be obtained from various datasets, while the RECIST marks in the current hospital picture archiving and communication systems (PACSs) can be employed for quick generation of large-scale and high-quality soft masks in particular.

C. Detailed Representation in Medical Images

High-resolution medical images are usually preferred in clinical practice due to more clear image structure and texture details, as well as the benefits to subsequent analysis and processing [32]. These detailed representations in high-resolution images are very important for accurate segmentation, especially in edge regions. Unfortunately, the tiny lung nodules are difficult to overcome the challenges of hardware, physical, and physiological factors to obtain high-resolution images under the existing imaging system. In recent years, some deep learning-based natural images or medical images SR methods [16], [33], [34], [35] have excellent performance in image resolution improvement. However, the lack of high-resolution training data limits the application of these methods. In addition, the potential of combining these methods with medical segmentation has not been fully exploited. In this work, we conduct self-supervised training on an SR model for medical images and smartly transfer the detailed representation knowledge in the trained SR model to the lung nodule segmentation pipeline.

III. METHODOLOGY

In this section, we propose a complete paradigm for accurate lung nodule segmentation. First, we introduce the soft mask and develop a pipeline to label it automatically. Then, we design a novel DSNet with detailed representation transfer and soft mask-based adversarial training framework.

A. Soft Mask for Lung Nodule Segmentation

1) Definition of Soft Mask: Lung nodules are usually labeled and segmented in the form of a binary mask in most methods. However, the binary mask has many disadvantages in accurately segmenting lung nodules. First, it cannot clearly describe the edge and morphological details of the lung nodules, which may cause anatomical information in these marginal regions to be ignored. Second, the labeling is extremely unbalanced, since most of the regions are labeled as nonlesion in the binary mask [30], and this imbalance impairs the training of the segmentation network. Third, the binary mask has the risk of overfitting in training and leads to poor generalization of the model.

Fig. 2. (a) Lung nodules with RECIST marks. The red line is the long axis, and the blue line is the short axis. (b) Trimaps. White pixels mean the lesion region. Black pixels mean the background region. Gray pixels mean the uncertain region. (c) Soft masks. (d) Binarized soft masks. The soft mask preserves edge details of lung nodules effectively.

To overcome the weakness of the binary mask, we introduce a new form of accurate labeling called soft mask for lung nodule segmentation, inspired by the image matting task [36]. We define the $i$th pixel of soft mask $M_i$ as a linear combination of the lesion label $L_i$ (the value is 1) and nonlesion background label $B_i$ (the value is 0)

$$M_i = \alpha_i L_i + (1 - \alpha_i) B_i$$  \hspace{1cm} (1)\\

where $\alpha$ means the extent of belonging to the lung nodules region. Here, the close-form-matting algorithm [37] is employed to solve the $\alpha$ matrix, which converts the problem into a closed-form solution through linear assumptions and the labeling pixels. The labeling pixels are the key prior information for solving the above problems and are represented by trimap, which consists of lesion pixels, background pixels, and uncertain area pixels shown in Fig. 2(b). Therefore, it is important to obtain a high-quality trimap.

2) Labeling of Soft Mask: We develop an automatic soft mask labeling pipeline to deal with priors from different datasets.

First, we initialize these different priors into trimaps composed of lesion region, background region, and the uncertain region, as shown in Fig. 2(b). Specifically, we design the following three strategies for the trimap initialization.

1) Trimap Generation With the Normal Binary Mask: We use morphological operations to process the binary mask. Specifically, the lesion region is obtained by the erosion operation, while the nonlesion background region is obtained by the dilation operation, and the uncertain region is in the middle.

2) Trimap Generation With the Binary Mask of Different Doctors: In order to improve the quality and reliability of labeling, some datasets are repeatedly labeled by different doctors. In the Lung Image Database Consortium (LIDC) dataset [38], the mask for each lung nodule is marked by four different doctors, as shown in Fig. 1(b). We set the intersection of these different masks labeled by doctors as the lesion region, while we set the complement of the union of these masks to the background region, and other pixels belong to the uncertain region. Formally,
it can be expressed as follows:
\[
\begin{align*}
L &= M_{d1} \cap M_{d2} \cap M_{d3} \cap M_{d4} \\
B &= M_{d1} \cup M_{d2} \cup M_{d3} \cup M_{d4} \\
U &= L \cup B
\end{align*}
\]
where \(L\) denotes lesion region, \(B\) denotes background region, and \(U\) denotes uncertain region, with \(M_{di}\) representing the manual annotation of the \(i\)th doctor.

3) Trimap Generation With RECIST Marks: RECIST marks [8] are commonly found in current hospital PACSs despite their coarseness. This means that converting these massive RECIST marks data into accurate masks will have great application value and potential. As shown in Fig. 2(a), RECIST marks have a long axis and a short axis to mark the diameters of the lesion. First, we use the regions marked on the long axis and the short axis as the prior information of the GrabCut [9] to obtain the initial rough binary masks. Then, we apply the morphological processing mentioned in strategy (1) to get the trimap.

After obtaining the reliable trimap, we use the close-form-matting algorithm [37] to generate soft masks shown in Fig. 2(c). Since solving \(\alpha\) for each pixel in (1) is an ill-posed problem, appropriate conditional assumptions need to be made. The close-form-matting algorithm [37] assumes that the foreground \(L\) and background \(B\) are almost unchanged in a small window centered on any pixel, i.e., the foreground and background satisfy local smoothness. Base on this assumption, (1) can be reexpressed as follows:
\[
\alpha_i \approx a M_i + b \quad \forall \ i \in w
\]
where \(a = 1/(L - B), \ B = -B/(F - B), \) and \(w\) is a small window in the image. This approximate expression suggests finding \(a, \alpha, \) and \(b\) that minimize the cost function as follows:
\[
J(\alpha, a, b) = \sum_{i \in M} \left( \sum_{j \in w_j} (\alpha_i - a_j M_i - b_j)^2 + \epsilon a_i^2 \right)
\]
where \(w_j\) is a small window around the pixel \(j\) and \(\epsilon a_i^2\) is a regularization term on \(a\) to obtain numerically stable and smooth solutions. For an image with \(N\) pixels, there are \(3N\) unknowns in (6). After a series of derivations and proofs [37], the cost function can be simplified to
\[
J(\alpha) = \alpha^T L \alpha
\]
where \(L\) is an \(N \times N\) matrix, whose \((i, j)\) entry is
\[
\sum_{k \mid (i, j) \in w_k} \left( \delta_{ij} - \frac{1}{|w_k|} \left( 1 + \frac{1}{|w_k|} + \frac{1}{\sigma_k^2} (M_i - \mu_k) (M_j - \mu_k) \right) \right)
\]
where \(\delta_{ij}\) is the Kronecker delta, and \(\mu_k\) and \(\sigma_k^2\) are the mean and variance of the intensities in the window \(w_k\) around \(k\), while \(|w_k|\) is the number of pixels in this window. In general, the close-form-matting algorithm constructs a Laplacian weight matrix by assuming the smoothness of the local small window and exploiting the correlation between each pixel in the local small window. Then, we take the explicit \(\alpha\) given by trimap as a constraint, and the optimal value of the cost function can be obtained by solving the minimum eigenvector of the \(L\) weight matrix, so as to obtain the solution of the entire \(\alpha\) matrix. Finally, substitute \(\alpha\) into (1) to obtain the soft mask.

Based on the above semiautomated pipeline, we label the lung nodules in the DeepLesion datasets [39] with RECIST marks to get soft masks. Then, we further binarize (with 0.5 as the threshold) soft masks to get an accurate binary mask [Fig. 2(d)] for the lung nodules segmentation task. Besides, the entire pipeline is efficient, which only takes 20 ms (for a single 256 \(\times\) 256 image: trimap 5 ms and close-form-matting 15 ms) for us to obtain a high-quality soft mask label. We label the lung nodules in the DeepLesion datasets [39] with RECIST marks and form a new dataset containing 1500 lung nodules named soft mask dataset of lung nodules soft mask dataset (LNSM).

B. Lung Nodule Segmentation Network With Detailed Representation Transfer and Soft Mask Supervision (DSNet)

Our DSNet can transform low-quality lung nodule images into a high-quality mask for accurate segmentation, as shown in Fig. 3. Specifically, we first design a DRTM to collect feature maps with detailed representations as the input to the network. Then, we employ a special backbone network to extract features and output prediction results.

1) Backbone Architecture: Based on the classic encoder–decoder architecture of U-Net [23], we design a special backbone to adapt to the accurate segmentation of lung nodules. As shown in Fig. 3, our backbone consists of basic blocks, upsampling blocks, and downsampling blocks. The convolutional block attention module (CBAM) block [40] is employed as the basic block of our backbone shown in Fig. 4, which uses both channel attention and spatial attention to enhance the expressive ability of feature maps and, thus, stimulate the performance of segmentation. In addition, the shortcut connection of ResNet [41] is also utilized to ensure the propagation of detailed information and gradients. In order to reduce the loss of detail caused by downsampling on the input image whose resolution is usually low, we use 4 \(\times\) 4 convolution layer with stride = 2, padding = 1, and rectified linear unit (ReLU) activation function as the downsampling block instead of pooling. The upsampling block of the decoder consists of 4 \(\times\) 4 transposed convolution and ReLU activation function, while the skip connection of U-Net is still retained to recover the detailed information. Noting that the design of the network architecture is not the focus of this work, so we employ some existing techniques and ideas to build the backbone network.

2) Segmentation Head: After a 3 \(\times\) 3 convolutional layer, the predicted segmentation mask is generated and accepts the joint supervision of the binary mask (after sigmoid) and the soft mask (after tanh, which normalizes the result to \([-1, 1]\]). We also normalize the ground truth of soft mask to the same range via the normalization function in PyTorch [42].

3) Detailed Representation Transfer: Lung nodules are usually very small (5–10 mm), so common CT systems can only obtain images of lung nodules with low resolution. The
Fig. 3. Overview of our DSNet. Our DSNet is composed of DRTM, backbone network, and adversarial training framework with soft mask.

Fig. 4. Details of the CBAM block in Fig. 3. Each $3 \times 3$ convolutional layer is followed by a ReLU activation function and a batch normalization layer.

Low-resolution image means the loss of detailed information, which is challenging for accurate segmentation of lung nodules with complex shapes and blurred edges. Inspired by SR methods, we reconstruct the detailed information in the high-resolution image and use it to stimulate the performance of accurate segmentation. Next, we will show how to train an SR model without high-resolution lung nodules data and apply the detailed information from the off-the-shelf SR model to the segmentation pipeline.

a) Self-supervised SR model: The lack of high-resolution lung nodules images makes it difficult to train an SR model. Due to the robustness of the SR methods [16] and the inspiration of the self-supervised pretraining methods [15], we self-supervised train an SR module with a general CT images dataset DeepLesion [39]. The DeepLesion dataset consists of 32,735 important clinical radiological findings (lesions, tumors, lymph nodes, and so on), and it can be universally adapted to the data of different organs. Specifically, we downsample the CT images of the DeepLesion dataset from $512 \times 512$ to $128 \times 128$ for training an SRGAN model [16] (from 128 to 512). Then, the trained off-the-shelf SR model is used as a teacher model to transfer the learned detailed knowledge to the DRTM for further lung nodules segmentation.

For transferring the detailed representation into the segmentation pipeline, we designed our DRTM to capture detailed representations as follows.

b) Detailed representation transfer module: First, our DRTM uses the context branch and the detail branch to extract feature maps, respectively, as shown in Fig. 5. For the context branch, we combine the interpolated low-resolution image and the output of the SR model and use the convolutional layer to extract the feature maps $F_c$. For the detail branch, we explicitly mine detailed representations from the feature map output of the last block of the SR model decoder. Specifically, given the input feature maps from SR model $F_{SR} \in \mathbb{R}^{H \times W \times C}$, we use two groups of convolutions to extract features from different receptive fields. We explicitly model high-frequency detail information by subtracting the outputs of these two convolutional layers as follows:

$$F_h = \text{sigmoid}(C_h(F_{SR}) - C_l(F_{SR}))$$

(9)

where $C_h$ denotes $1 \times 1$ pointwise convolution and $C_l$ denotes $3 \times 3$ dilated convolution with dilated rate 2. Here, we use pointwise convolution $C_h$ to preserve local high-frequency details and employ dilated convolution $C_l$ with a large receptive field to extract low-frequency semantic structure information. Therefore, the high-contrast pixels obtained by their subtraction are regarded as high-frequency detail information.

Then, $F_h$ is utilized to weight the original feature map to achieve the detailed texture enhancement as follows:

$$F_d = F_{SR} \times F_h$$

(10)
where $F_d$ is the feature maps output by the detail branch.

However, not all detailed representations from the SR model are conducive to the segmentation, and some noises are also introduced, as shown in Fig. 6(b). To suppress the noise and better fuse the detailed representation, we further propose a novel channelwise selective fusion (CSF) operation. Specifically, two compact vectors $v_c, v_d \in \mathbb{R}^{1 \times C}$ are created to enable the guidance for the precise and adaptive selections. This is achieved by the simple global average pooling (GAP) and fully connected (FC) layer

$$v_z = \text{GAP}(\text{Concat}(F_d, F_c))$$

where Concat denotes merging two feature maps along the channels. Then, we use GAP to generate the vector $v_z \in \mathbb{R}^{1 \times 2C}$

$$v_c, v_d = \text{FC}(v_z).$$

Here, the FC layer reduces $v_z$ to $\mathbb{R}^{1 \times C}$ dimension and then predicts two weight vectors $v_c$ and $v_d$ for the two branches, respectively.

Then, a soft attention across channels is used to select different spatial scales of information adaptively, which is guided by the compact vectors $v_c$ and $v_d$. Specifically, a softmax operator is applied on the channelwise digits

$$a = \frac{e^{v_c}}{e^{v_c} + e^{v_d}}, \quad b = \frac{e^{v_d}}{e^{v_c} + e^{v_d}}$$

where $a, b \in \mathbb{R}^{1 \times C}$ denote the soft attention vectors for $F_c$ and $F_d$, respectively. The final output feature maps $F_{\text{out}} \in \mathbb{R}^{H \times W \times C}$ are obtained through the weighted fusion

$$F_{\text{out}} = a \cdot F_c + b \cdot F_d, \quad a + b = 1.$$  

In addition, Fig. 6(c) and (d) shows the spatial attention maps in the CBAM block of DSNet. Here, Fig. 6(c) is to directly use the feature maps $F_{\text{SR}}$ of the SR model as input, while Fig. 6(d) is to use DRTM for representation transfer. We can see more impressive edge details [thanks to detail texture enhancement (DTM)] and less sharpening noises [thanks to CSF] with the help of our DRTM.

**C. Adversarial Training Framework With Soft Mask**

We use Sørensen–Dice coefficient (DICE) loss [43] as the segmentation loss for the supervision of ground-truth binary masks

$$\mathcal{L}_\text{Seg} = 1 - \frac{2 \sum_{i=1}^{N} (x_i y_i)}{\sum_{i=1}^{N} x_i^2 + \sum_{i=1}^{N} y_i^2}$$

where $y_i$ is the ground-truth segmentation mask $Y_{\text{bin}}$ for a given pixel $i$ and $x_i$ is the corresponding value in the predicted binary mask $X_{\text{bin}}$.

Soft mask loss is given to make the predicted soft mask has the same distribution as the ground-truth soft mask

$$\mathcal{L}_\text{Soft} = \|X_{\text{soft}} - Y_{\text{soft}}\|_{L1}$$

where $X_{\text{soft}}$ and $Y_{\text{soft}}$ denote ground-truth soft mask and predicted soft mask, respectively, and $\| \cdot \|_{L1}$ denotes the $L1$ distance.

Generative adversarial network (GAN) loss is introduced to further improve the quality of the predicted soft mask by conditional generative adversarial supervision [44]. As shown in Fig. 7, a pixelwise discriminator is designed to provide adversarial GAN loss for DSNet. Notice that pixel-level
discrimination is utilized to obtain an accurate pixelwise adversarial loss. Specifically, the overall objective can be expressed as follows:

$$G^* = \arg \min_S \max_D \mathcal{L}_{GAN}(S, D)$$

(17)

$$\mathcal{L}_{GAN} = \log D(x, y) + \log(1 - D(x, S(x)))$$

(18)

where our DSNet $S$ tries to minimize this objective against an adversarial discriminator $D$ that tries to maximize it. The $x$ denotes the lung nodules image (condition), $y$ denotes the real soft mask (ground truth), and $S(x)$ denotes the fake soft mask (predicted soft mask). In the training framework, our DSNet and the discriminator have trained alternately.

**D. Implementation**

The total loss function of our DSNet can be defined as follows:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{Seg} + \lambda_2 \mathcal{L}_{Soft} + \lambda_3 \mathcal{L}_{GAN}$$

(19)

where $\lambda_1 - \lambda_3$ are empirically set to 0.5, 100, and 1. The Adam optimizer [45] with poly learning rate policy is used to optimize the network with the training batch size set to 4. The learning rate decays from $2e^{-4}$, and the whole training process for all datasets typically converges in about 100 epochs with a single NVIDIA Titan V GPU.

**IV. EXPERIMENTS**

**A. Accurate Segmentation for Lung Nodules**

In this section, we have conducted comprehensive experiments on the proposed LNSM and LIDC [38] datasets to verify the effectiveness of our method on lung nodules segmentation. Following previous works [11], [14], DICE similarity (DICE), sensitivity (SEN), and positive predictive value (PPV) are used as the evaluation metrics. Here, DICE score is a region-level similarity measure that mainly focuses on the internal structural consistency of segmented objects, while SEN and PPV are pixel-level evaluation measures that equally consider the influence of each pixel to give the error value between the predicted mask and the ground truth. For convenient, we use the following denotations: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Therefore, these metrics are correspondingly defined as follows:

$$\text{DICE} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

(20)

$$\text{SEN} = \frac{TP}{TP + FN}$$

(21)

$$\text{PPV} = \frac{TP}{TP + FP}$$

(22)

1) **Comparison on Our LNSM Dataset**: Lung nodule images with RECIST marks in the DeepLesion dataset [39] are used to build our LNSM dataset. According to the strategy (3) proposed in Section III-A1, we obtained 1500 cropped $64 \times 64$ lung nodules images with soft masks and binary masks for training supervision. We show in Fig. 8 the steps to automatically obtain the soft mask via the RECIST marks. We randomly divide the dataset into training sets (1000 nodules), test sets (400 nodules), and validation sets (100 nodules), while all methods use the same dataset division. In the test, all methods utilize the same binary mask in LNSM as the ground truth to calculate metrics for a fair comparison.

Table I presents a quantitative comparison of some advanced methods on the LNSM dataset with the same dataset settings. The outputs are in “mean ± standard deviation” format. Specifically, U-Net [23], central focused CNN (CF-CNN) [11], dual-branch residual network
TABLE II

| Methods        | DICE (%) | SEN (%) | PPV (%) |
|----------------|----------|---------|---------|
| U-Net [23]     | 77.84 ± 21.7 | 77.98 ± 24.5 | 82.52 ± 21.5 |
| CF-CNN [11]    | 78.55 ± 12.5 | 86.01 ± 15.2 | 75.79 ± 14.7 |
| DB-ResNet [14] | 82.74 ± 10.2 | 89.05 ± 11.8 | 79.64 ± 13.5 |
| UNet++ [24]    | 80.54 ± 12.9 | 87.96 ± 17.2 | 79.18 ± 15.1 |
| INFNet [25]    | 81.01 ± 14.1 | 88.33 ± 13.7 | 75.58 ± 17.6 |
| INFNet† [25]   | 81.67 ± 13.7 | 87.25 ± 10.5 | 76.24 ± 13.3 |
| MedT† [28]     | 81.34 ± 7.9  | 87.96 ± 11.4 | 80.14 ± 15.3 |
| SoftGAN [17]   | 83.21 ± 7.0  | 89.14 ± 11.4 | 83.16 ± 12.1 |
| DSNet (our)    | 84.89 ± 7.2  | 90.56 ± 12.0 | 84.69 ± 13.4 |

(DB-ResNet) [14], UNet++ [24], lung infection segmentation network (INFNet) [25], soft mask-supervised GAN (SoftGAN) [17], and recent MedT [28] are compared. Here, † represents using SR image output by SR model we got in Section III-B3a as the input. We can see that our DSNet exceeds these state-of-the-art methods with a large margin, especially on the DICE and PPV metric scores.

2) Comparison on LIDC: We used a public lung nodules CT dataset from the LIDC and Image Database Resource Initiative [38] for further comparison. In this study, we studied 986 nodule samples annotated by four radiologists. Due to the differences in labeling between the four radiologists, the 50% consensus criterion [19] was used to generate the ground-truth binary masks. We use the method (strategy(2) in Section III-A1) of labeling soft mask to obtain ground-truth soft masks for training supervision. Then, we randomly partition these nodules into three subsets for training, validation, and testing with the number of nodules contained in each subset being 387, 55, and 544, respectively.

We compared our DSNet with some advanced segmentation methods, including U-Net [23], CF-CNN [11], DB-ResNet [14], UNet++ [24], INFNet [25], SoftGAN [17], and MedT [28], which are illustrated in Table II. Noted that † represents using SR image output by SR model we got in Section III-B3a as the input. The results show that our method outperforms other methods in all metrics. Compared with the competitive MedT†, our method leads 2.81%, 2.23%, and 3.68%, respectively. Notice that the performance of our DSNet is also significantly higher than that of our previous SoftGAN (84.89 versus 83.21 on DICE). Besides, the visual comparison between our DSNet and these methods on the LIDC dataset is shown in Fig. 9. It can be observed that our method has better contour and edge details.

B. Ablation Study

To validate the effectiveness of our modules and strategies, we conduct the following ablation studies on both LNSM and LIDC datasets.

1) Comprehensive Ablation Study: The ablation study is reported in Table III. We use the classic U-Net [23] trained with DICE loss as the baseline. After replacing with our backbone, the performance is significantly improved, which implies that the network structure is still crucial to lung nodules segmentation. Just using SR images as the input, the DICE score has also been slightly improved, indicating that high-resolution images can promote the accurate segmentation of lung nodules. Note that our DRTM further improves the performance indicating the importance of DTM and CSF. Finally, our DSNet achieves the best performance after applying complete adversarial training with soft mask supervision.

2) Ablation Study on Soft Mask Supervision: In addition, Table IV reports the comparison between our soft mask and other soft labels [30], [31]. Specifically, we use them to supervise the training of our DSNet under the same conditions, respectively. The results demonstrate that our soft mask is significantly better than other soft labels, due to the more accurate edge representation of our soft mask, which can better reduce the impact of imprecise boundary annotation.

3) Ablation Study on Detailed Representation Transfer: We also perform a meticulous ablation study for each component in our DRTM, as shown in Table V. The results show that the
Fig. 9. Visual comparison between our DSNet and advanced methods on the LIDC [38] dataset. Because of the soft mask supervision, our DSNet has an impressive visual effects that even exceeds binary ground truth.

Fig. 10. Visual comparison results between our DSNet and doctors. Our DSNet has an amazing visual effects that even exceeds binary ground truth (50% consensus).

C. Robustness Study

To further verify the accurate segmentation performance and generalization of our method, we design an ambitious and challenging experiment: compare our DSNet with human doctors. Specifically, we compare our DSNet trained on the LNSM dataset with four radiologists on LIDC. Note that our DSNet is not retrained, which means that the training set and the test set are cross domain.

Table VI shows the comparison results of our DSNet and four human doctors on the LIDC dataset. The DICE between any two masks (row and column) is reported in the table.

Due to the differences and contradictions in labeling between the four radiologists, the 50% consensus criterion [19] is often used as the ground truth. It can be observed from Table VI that our DSNet, even though it is trained on LNSM dataset, is still closer to 50% consensus than all doctors. On the one hand, the results demonstrate that our DSNet has strong robustness and can deal with cross-domain challenges, which implies the potential for clinical application. On the other hand, the results mean that the LNSM dataset we proposed is also universal and robust. In addition, Fig. 10 shows the visual comparison results between our DSNet and doctors. Our DSNet not only has a higher DICE score, but also has an impressive visual result that even exceeds ground truth (50% consensus). It is worth mentioning that our visual results also have been approved by clinical experts.
Fig. 11. Some visual results on ISBI and PROMISE. From left to right are the original image, the ground-truth binary mask of the dataset, the trimap image [obtained by labeling strategy (1)], and the soft mask and the results of our DSNet prediction.

D. Generalization Study

To verify the generalization of our method to other medical tasks, we perform experiments on International Symposium on Biomedical Imaging (ISBI) [46] and PROMISE [47]. The ISBI is a neuronal structure segmentation dataset, and the PROMISE is a prostate MRI dataset. After data augmentation, ISBI consists of three subsets: training, verification, and testing (210, 60, and 30 samples, respectively), while PROMISE consists of three subsets: training, verification, and testing (460, 65, and 130 samples, respectively).

We use the soft mask labeling strategy (1) proposed in Section III-A1 to obtain the soft masks for supervision. As shown in Table VII, our DSNet achieves the highest DICE and PPV scores and competitive SEN scores on these tasks that also require accurate segmentation. As shown in Fig. 11, our DSNet is also competent for medical segmentation tasks from cells (ISBI) to organs (PROMISE) and shows excellent generalization. Furthermore, it can be observed that the lead of our method on the ISBI and PROMISE datasets is not as pronounced as for the lung nodule segmentation task, due to two reasons: 1) our detailed representation migration is trained using the CT dataset (DeepLesion [39]), while ISBI [46] and PROMISE [47] are light microscopy and MRI images, respectively, and thus, there is potential cross-domain challenge and 2) lung nodules have rougher and more irregular edges than cells and prostates, so lung nodules can benefit more from the soft mask.

To explore the performance of our soft mask on the transformer-based segmentation method, we conducted experiments on the LNSM and LIDC datasets using MedT [28]. As can be observed from Table VIII, the performance of MedT is significantly improved on both LNSM and LIDC datasets after imposing our proposed soft mask.

E. Verified the Superiority of Soft Mask in Pathological Radiomics

Radiomics has demonstrated the relevance of a variety of quantitative imaging biomarkers for various cancer types. Usually, benign lesions are smooth (regular), and malignant lesions are burrs that can be judged by doctors directly according to the margin, sphericity, and spiculation characteristics. After doctors selected, we set up a classification experiment of smoothness and burrs to verify the soft mask’s ability of shape expression. We selected 200 lung nodule images on the LNSM data and 60 on LIDC, with the ratio of smooth to burr is 1:1, respectively. We extracted 93 features using the professional tool pyradiomic [48] and then used common machine learning methods [decision trees, k-nearest neighbor (KNN), and support vector machine (SVM)] to classify images based on these features. The results are shown in Table IX.
TABLE IX
CLASSIFICATION ACCURACY AND AREA UNDER CURVE (AUC) OF THE RADIOMICS FEATURES OF SOFT MASK AND BINARY Mask Are Obtained by Using Common Machine Learning Methods

| Methods  | Binary Mask | Soft Mask |
|----------|-------------|-----------|
|          | ACC | AUC | ACC | AUC |
| Decision tree | 0.61 | 0.81 | 0.83 | 0.81 |
| KNN       | 0.77 | 0.87 | 0.87 | 0.94 |
| SVM       | 0.84 | 0.90 | 0.88 | 0.91 |
| Decision tree | 0.68 | 0.63 | 0.83 | 0.88 |
| KNN       | 0.65 | 0.75 | 0.83 | 0.92 |
| SVM       | 0.72 | 0.76 | 0.83 | 0.92 |

and their intermediate region ROI can be used at the same time in radiomics analysis.

V. DISCUSSIONS

A. Time Consumption

One limitation of our method is that our method may introduce additional time consumption. On the one hand, the training phase requires additional time consumption when constructing the soft mask supervision. However, our annotation pipeline is automatic and efficient, as reported in Section III-A2. Our method only takes 20 ms (trimap 5 ms and close-form-matting 15 ms) for a 256 × 256 image to obtain a high-quality soft mask label. Once the soft mask dataset is obtained, this part of the time consumption will no longer be burdened. Besides, the discriminator is deprecated during the inference phase without increasing computational cost. On the other hand, our detailed representation transfer causes computational cost due to the SR model. But, note that the detailed representation transfer is designed to deal with low-resolution segmentation tasks, while the input size of the SR model is small, so the operating efficiency of the overall pipeline will not be significantly affected. Compared with the performance improvement, this extra calculation is worthy.

Table X shows our speed test results on the LNSM dataset.

| Config          | LNSM DICE(%) | LIDC DICE(%) | Inference time (ms) |
|-----------------|--------------|--------------|---------------------|
| w/o. DRF (SoGan [17]) | 91.63 ± 3.2 | 83.21 ± 7.0  | 5.1                 |
| w. DRF          | 93.77 ± 3.1  | 84.89 ± 7.2  | 6.7                 |

Deep learning algorithms based on neural networks have become popular in medical image processing, but traditional image processing algorithms can still bring good inspiration. In this work, we use some traditional image processing techniques and ideas, including close-form-matting [37], GrabCut [9], morphological processing [55], and high-frequency decomposition [56]. We explore and validate the possibility and potential of incorporating these traditional algorithms in a deep learning-based segmentation framework.

C. Relationship With Prompt Learning in NLP

Our soft mask supervision is similar to the prompt learning [54] that has attracted widespread attention in the natural language processing (NLP) field recently. It can be considered that the learning of the main task (lung nodules segmentation) is improved by adding a promotion task (soft mask generation). Therefore, our method can be regarded as an innovative attempt to prompt learning in the fields of computer vision and medical image processing.

D. Relationship With Traditional Image Processing Algorithms

Despite our soft mask yielding consistent performance improvements across various benchmarks, there is still a long road ahead before achieving practical application in clinical settings. We leave more effort to future work to improve the quality and reliability of soft mask.

VI. CONCLUSION AND FUTURE WORKS

In this work, we propose a complete solution for accurate lung nodule segmentation, including the soft mask labeling pipeline and a novel DSNet, which not only design the network structure but also the quality of the input image and ground truth. Our method is validated to be robust and universally effective on both lung nodule segmentation and other medical datasets and has great potential to segment other small lesions or small objects accurately while obtaining impressive visual results.

We consider the construction of a more suitable backbone network for soft mask supervision and detailed representation transfer as the future work, including the combination of
the popular visual transformer [57] and visual multi-layer perceptron (MLP) [58] with our proposed techniques. Besides, because the need for additional computation is a major limitation of our approach as discussed in Section V-A, we will explore further improvements in the operational efficiency of detailed representation transfer, such as introduce knowledge distillation techniques [59] and simplify the SR branch to achieve a balance of performance and efficiency.

REFERENCES

[1] C. Wang, P. W. Horby, F. G. Hayden, and G. F. Gao, “A novel coronavirus outbreak of global health concern,” Lancet, vol. 395, no. 10223, pp. 470–473, Feb. 2020.

[2] W. J. Kostis, A. P. Reeves, D. F. Yankelevitz, and C. I. Henschke, "Three-dimensional segmentation and growth-rate estimation of small pulmonary nodules in helical CT images," IEEE Trans. Med. Imag., vol. 22, no. 10, pp. 1259–1274, Oct. 2003.

[3] T. Kubota, A. K. Jerebko, M. Dewan, M. Salganicoff, and A. Krishnan, “Segmentation of pulmonary nodules of various densities with morphological approaches and convexity models," Med. Image Anal., vol. 15, no. 1, pp. 133–154, Feb. 2011.

[4] T. F. Chan and L. A. Vese, “Active contours without edges,” IEEE Trans. Image Process., vol. 10, no. 2, pp. 266–277, Feb. 2001.

[5] A. Levin, D. Lischinski, and Y. Weiss, “A closed-form solution to natural image matting,” in Proc. 16th Int. Conf. Comput. Vis. (ICCV). pp. 770–778.

[6] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “UNet++: A nested U-Net architecture for medical image segmentation,” in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (Lecture Notes in Computer Science), vol. 11045, D. Stoyanov et al., Eds., Granada, Spain: Springer, Sep. 2018, pp. 3–11, doi: 10.1007/978-3-030-00889-5_1.

[7] D. Wu et al., “Stratified learning of local anatomical context for lung nodules in CT images,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Jun. 2010, pp. 2791–2798.

[8] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation," in Proc. Int. Conf. Med. Image Comput. Assist. Intervent. Cham, Switzerland: Springer, 2015, pp. 234–241.

[9] E. Kats, J. Goldberger, and H. Greenspan, “Soft labeling by distilling anatomical knowledge for improved MS lesion segmentation," in Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI). pp. 1563–1566.

[10] J. Li, J. Goldberger, and R. Kankowski, “SoftSeg: Advantages of soft versus binary training for image segmentation," Med. Image Anal., vol. 32, no. 3, pp. 932–946, Mar. 2021.

[11] A. Levin, D. Lischinski, and Y. Weiss, “A closed-form solution to natural image matting," in Proc. 16th Int. Conf. Comput. Vis. (ICCV). pp. 770–778.

[12] C.-H. Pham et al., “Multiscale brain MRI super-resolution using deep 3D convolutional networks," in Proc. Med. Image Anal., vol. 34, no. 2, pp. 172–183, Aug. 2017.

[13] S. You et al., “Mild transformer: Gated axial-attention for medical image segmentation," in Proc. CVPR. pp. 770–778.

[14] C. Rother, V. Kolmogorov, and A. Blake, “‘GrabCut’ interactive foreground extraction using iterated graph cuts," ACM Trans. Graph., vol. 23, no. 3, pp. 309–314, 2004.

[15] S. Wang et al., “Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation," Med. Image Anal., vol. 40, pp. 172–183, Nov. 2017.

[16] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation," in Proc. Int. Conf. Med. Image Comput. Assist. Intervent. Cham, Switzerland: Springer, 2015, pp. 234–241.

[17] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation," in Proc. Int. Conf. Med. Image Comput. Assist. Intervent. Cham, Switzerland: Springer, 2015, pp. 234–241.

[18] C. Wang et al., “SoftGAN Towards accurate lung nodule segmentation via soft mask supervision," in Proc. Int. Conf. Multimedia Expo (ICME), Taiwan, Jul. 2022, pp. 1–6, doi: 10.1109/ICME52920.2022.9859677.
[42] A. Paszke et al., “PyTorch: An imperative style, high-performance deep learning library,” in Proc. Adv. Neural Inf. Process. Syst., vol. 32, 2019, pp. 1–12.

[43] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-Net: Fully convolutional neural networks for volumetric medical image segmentation,” in Proc. 4th Int. Conf. 3D Vis. (3DV), Oct. 2016, pp. 565–571.

[44] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” 2014, arXiv:1411.1784.

[45] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.

[46] G. Litjens et al., “Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge,” Med. Image Anal., vol. 18, no. 2, pp. 359–373, Feb. 2014.

[47] A. Cardona et al., “An integrated micro- and macroarchitectural analysis of the Drosophila brain by computer-assisted serial section electron microscopy,” PLoS Biol., vol. 8, no. 10, Oct. 2010, Art. no. e1000502.

[48] Pyradiomics. Pyradiomics Tool. Accessed: Sep. 27, 2023. [Online]. Available: https://pyradiomics.readthedocs.io/en/latest/index.html

[49] Y. Fu et al., “Elastic knowledge distillation by learning from reconstruction,” IEEE Trans. Neural Netw. Learn. Syst., vol. 34, no. 5, pp. 2647–2658, May 2023.

[50] A. Shamsi et al., “An uncertainty-aware transfer learning-based framework for COVID-19 diagnosis,” IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 4, pp. 1408–1417, Apr. 2021.

[51] S. Li et al., “Distilling a powerful student model via online knowledge distillation,” IEEE Trans. Neural Netw. Learn. Syst., early access, Mar. 7, 2022. doi: 10.1109/TNNLS.2022.3152732.

[52] M. Zhu, J. Li, N. Wang, and X. Gao, “Knowledge distillation for face photo–sketch synthesis,” IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 3, pp. 893–906, Feb. 2022.

[53] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing,” 2021, arXiv:2107.13586.

[54] D. Mellouli, T. M. Hamdani, J. J. Sanchez-Medina, M. Ben Ayed, and A. M. Alimi, “ Morphological convolutional neural network architecture for digit recognition,” IEEE Trans. Neural Netw. Learn. Syst., vol. 30, no. 9, pp. 2876–2885, Sep. 2019.

[55] W. Xie, J. Lei, Y. Cui, Y. Li, and Q. Du, “Hyperspectral pansharpening with deep priors,” IEEE Trans. Neural Netw. Learn. Syst., vol. 31, no. 5, pp. 1529–1543, May 2020.

[56] K. Han et al., “A survey on vision transformer,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 1, pp. 87–110, Jan. 2023.

[57] I. O. Tolstikhin et al., “MLP-mixer: An all-MLP architecture for vision,” in Proc. Adv. Neural Inf. Process. Syst., vol. 34, 2021, pp. 24261–24272.

[58] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” 2015, arXiv:1503.02531.