Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company’s public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Joint segmentation and detection of COVID-19 via a sequential region generation network

Jipeng Wu, Haibo Xu, Shengchuan Zhang, Xi Li, Jie Chen, Jiawen Zheng, Yue Gao, Yonghong Tian, Yongsheng Liang, Rongrong Ji

Abstract

The fast pandemics of coronavirus disease (COVID-19) has led to a devastating influence on global public health. In order to treat the disease, medical imaging emerges as a useful tool for diagnosis. However, the computed tomography (CT) diagnosis of COVID-19 requires experts’ extensive clinical experience. Therefore, it is essential to achieve rapid and accurate segmentation and detection of COVID-19. This paper proposes a simple yet efficient and general-purpose network, called Sequential Region Generation Network (SRGNet), to jointly detect and segment the lesion areas of COVID-19. SRGNet can make full use of the supervised segmentation information and then outputs multi-scale segmentation predictions. Through this, high-quality lesion-areas suggestions can be generated on the predicted segmentation maps, reducing the diagnosis cost. Simultaneously, the detection results conversely refine the segmentation map by a post-processing procedure, which significantly improves the segmentation accuracy. The superiorities of our SRGNet over the state-of-the-art methods are validated through extensive experiments on the built COVID-19 database.

1. Introduction

As the serious pandemic, coronavirus disease 2019 (COVID-19) [1,2] has infected 56,898,415 people and caused 1,360,381 deaths, according to the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU) [3] (updated 19 November 2020). COVID-19 originates from the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which can be spread by infectious viruses through breathing, coughing, and sneezing, etc. Effective screening of infected patients is critical to the fight against COVID-19, as shown by 1) early diagnosis of COVID-19 can reduce the transmission of the virus by isolating infected patients early; 2) early treatment can also improve the chance of survival, and these two steps for confirmed patients to confirm that the virus has been cleared. One of the gold criteria for the diagnosis of COVID-19 is polymerase chain reaction (PCR), which measures the biosafety level of nucleic acid (RNA) of SARS-CoV-2 in respiratory specimens, e.g., nasopharyngeal swabs, oropharyngeal swabs, middle nasal turbinate. However, PCR tests are short supply, laborious, time-consuming, and require a complex manual process. Besides, computed tomography (CT) can quantitatively analyze the progress of COVID-19, while PCR can only determine the presence or absence of COVID-19.

Another way to screen COVID-19 is to perform a radiographic examination by analyzing chest images, e.g., X-rays, and CT. Recent studies [4–8] have found characteristic abnormalities exist in the chest radiographic images of infected patients. Some researchers [9,10] suggest that chest CT should be considered the primary tool for screening COVID-19 in endemic areas. Compared to X-rays, CT screening is widely used due to its merit and three-dimensional view of the lung. In recent studies [11,12], typical signs of infection can be observed in CT images, e.g., early ground-glass opacities (GGO) and late lung consolidation. Therefore, a qualitative assessment of the longitudinal changes in infection and CT scans can provide useful and essential information for combating COVID-19.
However, manually delineating lung infections is tedious and time-consuming. Besides, the radiologist’s annotation of the infection is a very subjective task influenced by personal bias and clinical experience. Therefore, computer-aided systems are expected for an automatic and accurate radiographic interpretation in less time.

Recently, deep learning-based technology [13] has been widely used in COVID-19 image processing for its powerful image feature extraction. A COVID-Net [8] is proposed to detect COVID-19 from chest X-ray images. Wang et al. [14] adopted a deep learning method to extract COVID-19s graphical features for a clinical diagnostic test. Joaquin [15] introduced a transfer learning with ResNet50 to detect COVID-19. Butt et al. [16] developed a location-attention oriented model to calculate the infection probability of COVID-19. Tang et al. [17] presented a random forest method to assess the severity based on quantitative features. Shi et al. [18] utilized an infection Size Aware Random Forest method (ISARF) to classify COVID-19. Shan+ et al. [19] devised a deep-learning-based method to segment and quantify the infection regions from CT scans. Though plenty of deep-learning-based methods have been proposed to provide assistance and diagnose COVID-19, few publications are available to segment and detect COVID-19 in CT scans.

We aim to integrate the deep convolutional neural network (CNN) into the segmentation [20,21] and detection of the COVID-19 chest CT images. Fig. 1 shows the COVID-19 infections of chest CT images, which contain ground-glass opacities in the early stage, pulmonary consolidation in the late stage, and mix areas of both. The qualitative evaluation of infection areas and longitudinal changes in CT scans could reward useful information, e.g. lesion location, lesion size, and critical condition, in fighting against COVID-19. However, due to the unique structure of infection regions, it is difficult to distinguish the boundaries of COVID-19, which makes the accurate segmentation of COVID-19 lesions difficult. Besides, the manual delineation of COVID-19 infections is tedious and time-consuming, often influenced by individual bias and clinical experience. In the annotation process, the boundaries of COVID-19 infected areas can be better visualized by adjusting different parameters of window breadth and window locations, as shown in Fig. 1, which is beneficial to the segmentation of COVID-19.

In this work, a light-weight yet accurate framework is presented, termed Sequential Region Generate Network (SRGNet), which jointly detects and segments the chest CT images of COVID-19 infections. SRGNet aims to fully mine the supervisory information of COVID-19 infection images. We first design a context enhancement module to increase the contextual information between encoder and decoder. Then, a novel edge loss is designed to augment the contextual information of the image. Thus, the bounding boxes of the segmented image across different layers can be generated to reduce the computation of the detection task and mitigate the problem of an unbalanced category in detecting COVID-19. Lastly, the predicted boxes are further fed forward to refine the segmentation results via a post-processing.

To summarize, the main contributions of this paper are listed below: 1) We propose a new framework, termed SRGNet, to jointly detect and segment COVID-19 infection images in a mutually reinforcing manner. 2) A novel edge auxiliary loss function is designed to enhance the spatial constraints of features. 3) We enhance the contextual information by context enhancement modules from low-level layers to high-level layers.

2. Related works

In this section, we introduce the three types of methods related to our work, including Semantic Segmentation, Object detection and Multi-tasks learning.
**Semantic Segmentation:** A fully convolutional network (FCN) was proposed in [26] for semantic segmentation, which classifies each pixel of the image. Following FCN, [27] introduces Atrous Spatial Pyramid Pooling (ASPP) with different rates and the dilated convolution, which increases the receptive field with a little parameter increasing. Zhao et al. [28] proposed pyramid pooling module, which contains different scale information between different sub-regions, and improves the ability to obtain contextual information. Badrinarayanan et al. [29] introduces a decoder network, along with a pixel-wise classification layer. The decoder’s primary function is to restore the feature map to the full input resolution at the pixel level. Another standard model is U-Net [30], which contains a contracting path to capture context and a symmetric expanding path that enables precise localization. Furthermore, [31] extended U-Net to leverage information in multi-modal data. Each available image modality is processed differently to exploit their unique information instead of combining them at the input. Song [32] proposed a novel generative model-based framework for CT image segmentation and proposed a novel pixel-region loss function to handle data unbalance problem. Fan et al. [33] proposed a lung infection segmentation deep network to automatically identify infected regions from chest CT scans, which uses implicit attention and explicit edge-attention to model the boundaries and enhance the representations. Boldsen et al. [34] utilized a machine learning algorithm with decision trees and spatial regularization to outline the lesion. Abulnaga and Rubin [35] introduced the pyramid pooling method to combine global and local contextual information to segment lesions from CT images. Zhang et al. [36] equipped 3D deep convolutional network to learn 3D contextual information efficiently. However, a 3D convolution network is computationally intensive and time-consuming to predict infection areas. Liu et al. [37] proposed a novel deep residual attention convolutional neural network to accurately and simultaneously segment and quantify ischemic stroke and WMH lesions in the MRI images. Wang et al. [38] introduced a hybrid loss function to pay more attention to lesion regions and high-level contextual consistency.

**Object Detection:** Object detection is a fundamental task in computer vision, which locates and classifies objects in an image. Generally, the object detection method can be divided into two categories: One-stage detection and Two-stage detection. One-stage methods, e.g., YOLO [39], SSD [40], DSSD [41], YOLO v2 [42], directly predict object class and location in a single network. Therefore, it has a faster detection speed than two-stage method. Two-stage methods generate a series of candidate region proposals as samples by the algorithm and then classifies the samples through the convolutional neural network. Among them, R-CNN [43], Fast R-CNN [44] and Faster R-CNN [45] are the most representative frameworks. R-CNN uses selective search to generate region proposals and then feed those region proposals into classification and regression network. Fast R-CNN feeds the image to a convolutional network and crops a set region proposal through ROI pooling layers to extract a fixed-size feature vector, leading to a faster speed and better accuracy. Faster R-CNN is a more flexible and robust deep learning framework, which further replaces regional recommendations with regional suggestion networks. With the continuous public of CT data, many variant algorithms for CT images have been proposed. For example, [46] presents a threshold-based method to separate the candidate nodules from other structures in CT images of the lungs. Xie et al. [47] adjusts Faster R-CNN structure with two region proposal networks and a deconvolutional layer to detect nodule candidates. Wang and Wong [8] leverages long range connections to computational complexity to reduce the computational complexity in COVID-19 detection.

**Multi-tasks Learning:** Recently, many multi-task learning frameworks for detection and segmentation have been proposed. For example, [22] implemented the task of segmentation and detection in the last layer of the encoder, which suggested multi-task learning could be effectively optimized for both tasks. Brazil et al. [22] implemented a segmentation infusion network to jointly supervise semantic segmentation and object detection in the last encoder layer, that the additional supervision helps guide features in shared layers to be more sophisticated in object detector. Mao et al. [23] proposed a novel hyperlearner network to learn the CNN based detection, which benefited from the extra feature. He et al. [24] used a different branch to parallel predict a segmentation mask, and the other existing branch is implemented for bounding box recognition. Dwornik et al. [25] jointly performed object detection and semantic segmentation in a forward pass that benefits from each other in terms of accuracy. All these frameworks of [22–25] are compared in Fig. 2. Unlike previous works, our SRGNet performs semantic segmentation in multiple layers in both encoder and decoder, then region proposals for detection are directly generated from the segment maps. Meanwhile, the previous predicted segmentation layer play as a partial input to the next predicted segmentation layer, and detected lesions refine segmentation result in post-processing.

3. The proposed method

The proposed method is a 2D patch-based deep learning framework to jointly detect and segment the lesion areas of COVID-19 from CT images. Firstly, the network architecture is introduced in Section 3.1. Secondly, the context enhancement module is presented in Section 3.2, which enhances the context information of each pixel in the image feature through multiple convolutions for the coexistence of lesions of different sizes. Thirdly, the loss functions of detection and segmentation are described in Section 3.3. Then, a random region proposal is shown in Section 3.4. Finally, a post-processing procedure is introduced in Section 3.5.

3.1. Model overview

Our key insight is that the real scene’s structure information is explicitly useful for reasoning lesions from a CT image, as the lesion is closely associated with the local structure and surface regions. The traditional deep-based-methods, e.g. U-Net [30], suffer from unsatisfactory performance since they do not adequately take advantage of structure information, e.g., the corners and edges.

We design an innovative hybrid model to fully explore the structure information at a single image’s multi-scale to solve this. The overview of the proposed SRGNet is shown in Fig. 3. Let X be an input of the network, where X is a CT slice image. Our goal is to train a CNN that can simultaneously locate and segment the lesion areas of COVID-19 from X. We first produce a set of segmented features S = (S₁, S₂, ..., Sₙ), which convolutes with the intermediate features (X₁, X₂, ..., Xₚ) by a context enhancement module. Here, Xᵢ is concatenated by the last layer of the ith block and previous segmented map Sᵢ₋₁. Thus, the semantic information can be spread through the second path network. Next, we further enhance the feature context information by using edge auxiliary loss, which benefits the prediction of segmentation and detection. Then, we directly obtain bounding boxes from the sequential series (S₁, S₂, ..., Sₙ), and a random region proposal method is explored to train the detection network. Finally, a post-processing procedure is proposed to refine the predicted segmentation results.

3.2. Context enhancement

A key component of SRGNet is the Context Enhancement (CE) module to enhance the contextual information of the inter-
Fig. 2. Different frameworks of joint object detection and semantic segmentation. (a) Both detection and segmentation are added to the last layer of the encoder [22]. (b) The segmentation branch is used to refine the detection branch [23]. (c) Each layer of the decoder detects the object in various scales and selects the ROI areas for segmentation [24]. (d) Each layer of the decoder detects the object, and the multi-scale features are combined for segmentation [25]. (e) Our proposed SRPNet, in which each layer of the encoder and decoder can be used for segmentation, and region proposals for detection are directly generated from the segment maps. Meanwhile, the previous predicted segmentation layer is as a partial input to the next predicted segmentation layer. Among them, “S” and “D” denote segmentation and detection prediction, respectively.

Fig. 3. The proposed SRPNet. All the encoder and decoder blocks are used for segmentation, and the bounding boxes are directly generated from the segmented maps. Meanwhile, the previous predicted segmentation layer will be considered a partial input to the next predicted segmentation layer. In this stage, we add a context enhancement module (CE) and edge loss (EL) to enhance features’ contextual information. Then, we increase the number of practical bounding boxes through random region proposals and extract responding features from the network’s last layer. Finally, ROI features can be directly resized from these features and as the input to train the detection network with classification loss and regression loss. The predicted bounding boxes can be further used to refine the predicted segmentation results by post-processing during the testing.

CE is used to enhance the local contextual information for each intermediate feature $X_i$. More specifically, CE consists of a $3 \times 3$ convolution, a $3 \times 3$ dilated convolution with dilation 2 and a $5 \times 5$ depth separable convolution to enhance the contextual information of lesion areas. As shown in Fig. 4, $3 \times 3$ convolution involves each point with 8 adjacent points, while $3 \times 3$ dilated convolution performs for convolution operation with adjacent point of larger intervals, which increase the receptive field of the network. Compared to $3 \times 3$ convolution, $5 \times 5$ convolution consumes about 2.78 times of computations in the same input channel and output channel. To this end, we select $5 \times 5$ depth separable convolution to embed into our module, which greatly reduces the computation of CE module. Among them, $X_i$ is the input of in the $r$th block, and $X_1, X_2$ and $X_3$ are the output features of $3 \times 3$ convolution, $3 \times 3$ dilated convolution and $5 \times 5$ depth separable convolution, respectively. To further reduce the computation cost, we decrease the output channel of $X_1$, $X_2$ and $X_3$. In details, the channel number of $X_1$ is half of $X_i$, and $X_2$ and $X_3$ are a quarter of $X_i$. Without loss of generality, we can assume that each position is more relevant to the nearest positions than other distant positions. Thus, we utilize more output channels for the $3 \times 3$ convolution. Then, we concatenate $X_1$, $X_2$ and $X_3$, and increase the non-linearity by Leaky ReLU. As a result, we obtain the enhanced contextual features $X_4$, and then concatenate $X_4$ and $X_i$ as the input $X_5$ of the CE module. The feature $X_5$ after multi-scale local enhancement has twice as many channels as $X_i$. Finally, we apply a $3 \times 3$ convolution to output the segmentation map $S_i$ with two channels.
3.3. Loss functions

Semantic Segmentation. In addition to supervising the pixel-level labels of the lesion area, we also add a multi-scale segmentation auxiliary loss to tune contextual information of each segmentation map from the encoder and decoder modules. The 7th stage uses the softmax cross-entropy loss and an edge loss, defined as:

$$L_s^r = -\frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{m=1}^{M} S_{r}(h, w, m) \log \frac{\sigma_{S}(h, w, m + \kappa L_r^c)}{\sum_{\sigma=1}^{C} \sigma_{S}(h, w, m + \kappa L_r^c)}, \quad (1)$$

where $S_r$ denotes the predicted feature map of size $H \times W \times M$, $S_r$ denotes the corresponding ground-truth, $M$ is the number of categories, and $\sigma$ is the index. $\kappa$ is used to balance the softmax cross entropy loss and the edge loss $L_r^c$.

The edge loss (EL) is another crucial component of SRGNet, enhancing the spatial constraints of features. We adopt $L_r^c$ as the auxiliary loss and set $\kappa$ to 0.02 to balance the weight between the pixel level and the spatial level. $L_r^c$ is defined according to the spatial edge maps and is calculated in the horizontal and vertical directions by introducing an offset parameter of $r$. The edge maps are defined as:

$$E_{r}^{m_{ho}}(x, y) = \begin{cases} 1, & \text{if } p_{r}^{k} \neq p_{r+1}^{k}, \\ 0, & \text{otherwise} \end{cases}$$

$$E_{r}^{m_{io}}(x, y) = \begin{cases} 1, & \text{if } q_{r}^{k} \neq q_{r+1}^{k}, \\ 0, & \text{otherwise} \end{cases}$$

$$\hat{E}_{r}^{m_{ho}}(x, y) = \begin{cases} 1, & \text{if } q_{r}^{k} \neq q_{r+1}^{k}, \\ 0, & \text{otherwise} \end{cases}$$

$$\hat{E}_{r}^{m_{io}}(x, y) = \begin{cases} 1, & \text{if } q_{r}^{k} \neq q_{r+1}^{k}, \\ 0, & \text{otherwise} \end{cases}$$

(2)

where $x$ and $y$ denote the $x$th row and $y$th column of the matrix. $m_{ho}$ and $m_{io}$ are short for horizontal and vertical, respectively. $p_{r}$ and $q_{r}$ can directly generate from $S$ and $S'$ through a argmax operation in channel $M$, respectively. $p_{r}^{k}$ and $q_{r}^{k}$ denote the category value of $p_{r}$ and $q_{r}$ at $(x, y)$, respectively. In our scenario, such a setting aims to capture the edges in the horizontal and vertical directions, which have a high activation on multiple steps. The final loss is calculated among the edge maps $E_{r}^{m_{ho}}, E_{r}^{m_{io}}, \hat{E}_{r}^{m_{ho}}$ and $\hat{E}_{r}^{m_{io}}$ by a normalized $L_1$ loss.

Based on the edge maps introduced above, we consider the following loss function to minimize the difference edge maps between $p_{r}$ (i.e., $E_{r}^{m_{ho}}$ and $E_{r}^{m_{io}}$) and $q_{r}$ (i.e., $\hat{E}_{r}^{m_{ho}}$ and $\hat{E}_{r}^{m_{io}}$):

$$L_r^c = \sum_{r=1}^{R} \phi_1 \left( \frac{1}{N_{ho}} \left\| E_{r}^{m_{ho}} - \hat{E}_{r}^{m_{ho}} \right\|_1 + \frac{1}{N_{io}} \left\| E_{r}^{m_{io}} - \hat{E}_{r}^{m_{io}} \right\|_1 \right). \quad (3)$$

where $N_{ho}$ and $N_{io}$ are the pixel numbers of different edge maps in the horizontal and vertical directions. In our experiment, $r$ is set to 3, and $\phi_1, \phi_2, \phi_4$ are set to 2, 1, and 0.5, respectively.

The total segmentation loss of all the stages is a weighted sum of $L_s^r$:

$$L_s = \sum_{t=1}^{N} \lambda_t L_s^r. \quad (4)$$

Here, we introduce the hyperparameters $\{\lambda_t\}$ to balance the main segmentation loss function $L_N$ and the auxiliary losses $\{L_r^c\}, 1 \leq t \leq N - 1$. $\lambda_t$ gradually increases as $t$ becomes larger. More details are discussed in the ablation study of the experiments.

Object Detection. To train the region proposal network of detection task, we assign a binary class label to each anchor, conducted by ROI features in the last layer. Thus, we assign positive labels to these anchors with Intersection-over-Union (IoU) higher than 0.7 and assign negative labels if IoU ratios were lower than 0.3. Note that anchors neither positive nor negative do not contribute to training the network. With these considerations, we minimize the following objective function:

$$L_D = \frac{1}{N_{cls}} \sum_{l=1}^{N_{cls}} L_{cls}(p_l, p_l^*) + \beta \frac{1}{N_{reg}} \sum_{l=1}^{N_{reg}} p_l^* L_{reg}(t_l, t_l^*), \quad (5)$$

where $l$ is the anchor index, $p_l$ and $p_l^*$ denote the predicted probability and ground-truth probability of anchor $l$, respectively. $p_l^*$ equals to 1 if the anchor is positive, otherwise 0. $t_l$ is the 4 parameter coordinate vector of a predicted bounding box and $t_l^*$ is the vector of the ground-truth bounding box associated with a positive anchor. The classification loss $L_{cls}$ is the two-class cross-entropy loss, and the regression loss $L_{reg}$ is the smooth $L_1$ loss, as defined in [44]. $p_l^* L_{reg}$ makes $L_{reg}$ be activated only when the anchor is positive ($p_l^* = 1$). The two terms $L_{cls}$ and $L_{reg}$ are weighted by a parameter $\beta$, which is set to 10. Note that $L_s$ and $L_D$ are trained...
separately. In this way, we first use the segmentation loss to learn the segmentation task. When the network converges, we keep the parameters of the segmentation task fixed and continue to use the detection loss to learn the parameters of the detection task.

3.4. Random region proposals

The second key component of SRGNet is the random region proposal method, which is shown in Fig. 5. The bounding boxes are directly extracted from the segmented image. In detail, each module in the network outputs a segmentation map, and the bounding boxes can directly be obtained from the segmented lesions, which is expressed as follows:

$$\{x_j, y_j, w_j, h_j\} = \text{Rect}(S_j),$$  

(6)

where $j \in J$, and $J$ is the number of $J$ bounding boxes in the $t$th segmentation image. Rect denotes the rectangle operation, which returns the location of bounding boxes for each lesion area. $x_j$ and $y_j$ denote the coordinates of the upper left corner of the $j$th bounding box in $t$th segmentation map. $w_j$ and $h_j$ denote the width and height of the $j$th bounding box in $t$th segmentation map, respectively.

However, those bounding boxes are not enough for training. Therefore, we use the random region proposal (RRP) method to increase the number of bounding boxes, which is defined as:

$$\{x_{jk}, y_{jk}, w_{jk}, h_{jk}\} = F_R(x_j, y_j, w_j, h_j)$$  

(7)

where $F_R$ is the function of random region proposal, which randomly changes the coordinates, heights, and widths of the bounding boxes within a certain scale train the detection model. In these experiments, we set the horizontal and vertical coordinates to be randomly shifted within 10 pixels. Moreover, the widths and heights to be randomly scaled between 0.8 and 1.2. $k$ indicates that each bounding box can be expanded to $k$ bounding boxes. $x_{jk}$, $y_{jk}$, $w_{jk}$, and $h_{jk}$ denote the location information of bounding boxes after expansion. For each segmented image $S_t$, we can get $K \times J$ bounding boxes. Finally, we select those bounding boxes as the anchor part, which effectively alleviates the unbalanced category problem when training the detection network.

3.5. Post-processing

In the test phase, we directly obtain bounding boxes from the last segmentation map. A certain number of anchor boxes are generated by a specific offset and scaling from these bounding boxes. Those anchor boxes are regressed and classified by a network, and 3 anchor boxes with the highest classification scores are used to calculate the overlapping rates with the bounding boxes from segmentation. If an overlapping rate is less than 20%, we consider that the bounding box’s detected lesion is invalid.

4. Experimental results

**Dataset.** Currently, there are few publicly available COVID-19 datasets with pixel-level markings. To this end, we constructed a new COVID-19 segmented (COVID) dataset. In this section, we introduce data collection, professional marking, Fig. 6 shows some examples of our COVID-19 dataset.

A. Data collection To protect the privacy of patients, we ignore their personal information in the dataset. We collected 313,167 CT slices from 438 patients, all of which are positive for COVID-19 and confirmed by RT-PCR tests. All images were acquired at 2 Chinese hospitals between January and February 2020. A medium sharp reconstruction algorithm reconstructed the CT slices with a thickness range from 0.5 mm to 1.25 mm. As in the previous study [48], we did not take patients with community-acquired pneumonia (CAP) into consideration.

B. Data annotation Although we have captured enough COVID-19 chest CT image data, accurate annotated labels are essential for deep learning. To this end, we formed a team of 3 annotators with deep radiology background and proficient annotation skills to annotate the areas and boundaries of the COVID-19 infected area, and a senior radiologist checks the final annotation with first-line clinical experience in COVID-19. For the segmentation task, we perform pixel-level labeling as strategies: 1. To save labeling time, the radiologists randomly select CT scan images of 50 patients. In this step, our goal is to label infected areas with pixel-level annotations. 2. To generate high-quality annotations, we invite a senior radiologist to refine the labeling marks for cross-validation. Some inaccurately labeled images have been removed in this stage.

After the above annotation procedures, we finally obtain 12,604 pixel-level labeled CT slices from 50 COVID-19 patients with a resolution of 512 × 512. Note that our detection labels can be directly acquired from corresponding segmentation labels, which is shown in Fig. 6. Figure 7 introduced the age and gender distribution of patients in our dataset. Among them, 1,360 slices were marked with lesions. To reduce the labeling process’s errors, we only select infected slices with lesion labeling for training and testing. We randomly split the dataset into 80% training and 20% testing on the patient-level, i.e., we select 1,024 CT slices for training and the other 336 for testing.

**Training Detail.** Our model is implemented with Tensorflow-GPU 1.8.0. All training and testing are done on a single Tesla V100 GPU using CUDA 10.0 and CUDNN 7.5. In detail, we train the network parameters over 120 epochs using the training set with a learning rate of 0.001. We utilize the stochastic gradient descent (SGD) with a momentum of 0.99 and a weight decay of 0.0005 in training for all experiments. The batch size is set to 12.

**Limitation.** Compared with 2D segmentation, 3D segmentation can better use spatial context information in the third dimension. However, when doctors mark covid data, some of them will inevitably be mismarked. If we use data for 3D segmentation, some incorrect annotations may be introduced. Moreover, the computation of 3D convolution is much higher than 2D convolution. Limited by GPU video memory, 3D convolution needs to use local features for training, and some context information will be lost. Therefore, we mainly focus on the 2D segmentation and detection of ischemic stroke.

4.1. Comparison to state-of-the-art methods

We compare SRGNet with state-of-the-art methods on the COVID-19 dataset. Table 1 shows the comparison results with Faster-RCNN [45], Cascade-RCNN [49], Scratch-Faster-RCNN [50], U-Net [30], U-Net++ [51], Attention U-Net [52], Mask-RCNN [24], BlitzNet [25], yolact [53] and Solo [54]. Codes of these methods are publicly available, and we follow the authors’ instructions to re-
train the models on the COVID-19 dataset. Note that Faster-RCNN [45], Cascade-RCNN [49] and Scratch-Faster-RCNN [50] detect the lesions of COVID-19, which are compared in terms of the accuracy $AP_{0.0.2.0.5}$; 0, 0.2 and 0.5 are threshold, comparing to IoU between predicted bounding box and corresponding ground truth, to judge whether the predicted bounding boxes is right or not. U-Net [30], U-Net++ [51], Attention U-Net [52], Mask-RCNN [24] and BlitzNet [25] are semantic segmentation methods, which are compared in terms of IoU and Dice [55], Mask-RCNN [24], BlitzNet [25], Yolact [53] and Solo [54] jointly detect and segment the lesions of COVID-19. In Table 1, "-" indicates that the method is not compared on this metric.

The evaluation result of SRGNet and other compared methods are displayed in Table 1. We can see that Cascade-RCNN, which uses the previous detection results as the latter detection stage input, outperforms Faster-RCNN, and Scratch-Faster-RCNN on average precision. It proves our conjecture that mining structure information in previous levels can improve the prediction of the network. Among the segmentation methods, Attention U-Net achieves the best performance, owing to Attention U-Net increases the network's receptive field, which contributes to the segmentation task. In multi-task learning, BlitzNet performs better than Mask-RCNN in five metrics, including pixel accuracy, IOU, Dice, $AP_{0.2}$, and $AP_{0.5}$. Yolact [53] directly generates the segmentation and detection result in one stage. Mask-RCNN first performs the detection task and then performs image segmentation in the detection box, which damages the segmentation task's contextual information. BlitzNet simultaneously predicts the segmentation and detection results in the decoder network, which further utilizes the supervision information than Mask-RCNN. Solo [54] outputs segmentation masks on whole image. However, the proposed SRGNet is more stable across various evaluation metrics. From Table 1, we can see that the multi-task methods, e.g., Mask-RCNN, BlitzNet, and our SRGNet, outperform the traditional lesion detection task, e.g., Faster-RCNN, Cascade-RCNN, and Scratch-Faster-RCNN. It proves that the extra supervision of segmentation can effectively improve the performance of lesion detection. In terms of AP, Dice, and IOU's best-known metric, SRGNet achieves the best performance, demonstrating the

Table 1

Segmentation and detection comparison among our SRGNet and other state-of-the-art methods on the COVID-19 dataset.

| Method                  | Pixel Acc | IoU | Dice | $AP_{0.2}$ | $AP_{0.5}$ |
|-------------------------|-----------|-----|------|------------|------------|
| Faster-RCNN [45]        | -         | -   | -    | 83.2       | 72.4       | 58.7 |
| Cascade-RCNN [49]       | -         | -   | -    | 79.9       | 72.4       | 58.7 |
| Scratch-Faster-RCNN [50]| -         | -   | -    | 85.4       | 74.2       | 64.3 |
| U-Net [30]              | 77.7      | 66.8| 78.9 |            |            |      |
| U-Net++ [51]            | 78.2      | 67.8| 80.2 |            |            |      |
| Attention U-Net [52]    | 81.4      | 69.7| 82.1 |            |            |      |
| Mask-RCNN [24]          | 78.5      | 67.2| 80.7 | 89.7       | 77.8       | 66.4 |
| BlitzNet [25]           | 81.7      | 69.5| 81.1 | 89.2       | 78.3       | 66.2 |
| Yolact [53]             | 69.6      | 58.4| 74.2 | 84.3       | 73.1       | 61.9 |
| Solo [54]               | 82.4      | 70.1| 83.0 | 90.6       | 78.1       | 66.8 |
| SRGNet (Ours)           | **83.2**  | **71.1**| **83.0**| **90.6**  | **78.7**  | **67.2** |

Fig. 6. Examples of our COVID-19 dataset, including CT scan images, segmentation labels and detection labels from top to bottom. Note detection labels are directly acquired from corresponding segmentation labels.

Fig. 7. The age and gender distribution of patients in our dataset.
superiority of SRGNet in COVID-19 infected area segmentation and detection.

Moreover, some visual comparisons among U-Net [30], Faster-RCNN [45], Yolact [53], and SRGNet are shown in Fig. 8. We can observe that the results of SRGNet are very close to the ground truth.

4.2. Ablation study

In this section, we first analyze the effects of different components in the proposed SRGNet. To examine the context enhancement (CE), edge loss (EL), random region proposals (RRP) and post-processing modules, we report the experimental results in Table 2. For the base network, U-Net based framework is adopted to our network, and the softmax cross-entropy loss is merely used for training. We use resnet50 as the encoder network and four modules on the decoder network. Each decoder module has 2 convolutional layers, including a $3 \times 3$ deconvolution with stride 2 and a $3 \times 3$ convolution with stride 1. Note $3 \times 3$ deconvolution is used to restore the resolution of the feature.

It can be seen that all three modules have a significant contribution to the excellent performance of SRGNet, especially the context enhancement and edge loss. Among them, “Base+CE (decoder)” only uses the context enhancement in the decoder network, while “Base+CE” applies it in both encoder and decoder, which proved the effectiveness of utilizing the feature from the whole network. From Table 2, SRGNet obtains 0.5%, 1.0% and 2.4% improvement in Dice, IoU, and Pixel accuracy when CE is applied in decoder modules, and further gets 0.2%, 0.5% and 0.5% improvement when CE adopted in both encoder and decoder modules. Simultaneously, benefited from the context enhancement module, there is also a corresponding improvement in detection. Then we use the edge auxiliary loss function to optimize the semantic spatial information of the whole network. In the same way as the context enhancement module, edge loss is applied to the feature with different resolutions among low-layers and high-layers to obtain the spatial optimization in different resolutions. In this way, we obtain 0.5%, 0.6%, and 1.8% improvement in Dice, IoU, and Pixel accuracy, respectively. Finally, SRGNet jointly detects and segments ischemic stroke by post-processing. Table 2 shows that our SRGNet obtain 0.2%, 0.3% and 0.2 improvement in Dice, IoU, and Pixel accuracy obtained the corresponding improvement in lesion detection. “SRGNet(without RRP)” means SRGNet does not use RRP. In this way, we randomly crop $16 \times 16$ features as the anchor of the detection task. As can be seen, we further obtain 0.6%, 1.6% and 3.3% improvement in $A_0$, $A_{0.2}$ and $A_{0.5}$ when RRP embedded with SRGNet.

Then, we explore hyper-parameters’ influence on the experimental results, including $\lambda_d$ and $\kappa$. As shown in Fig. 9, the first set of hyperparameters is set to 0.2, while the second set of hyperparameters increases in the segment, and the third set of hyperparameters increases linearly. From Fig. 9, the third set of hyperparameters yields the best result. It proved that later modules are more conducive to the prediction of the network. Fig. 10 show the segmentation result with different $\kappa$ settings. When $\kappa$ is set to 0.02, the network obtains the IOU and Dice’s best performance. Note none of Figs. 9 and 10 use post-processing.
5. Conclusion

We have proposed an effective Sequential Region Generation Network (SRGNet) to jointly detect and segment the lesions of COVID-19 in a mutually reinforcing manner. SRGNet directly generates bounding boxes from the segmented images in various layers, which decreases the computation cost and alleviates the unbalanced category problem when training the detection network. The predicted bounding boxes further refine the segmented result by a post-processing process. Besides, SRGNet employs a context enhancement module and edge loss to enhance the local and global information of the features from the low-level layers to the high-level layers, which further improves the accuracy of segmentation and detection of COVID-19. The experimental results have demonstrated the superior performance of SRGNet. However, there are still some limitations worth noting.

Firstly, the detection and segmentation of the COVID-19 lesion area are highly dependent on the annotation of the clinical experts, which is very time-consuming and affected by the prior knowledge of experts. Secondly, the affected area of COVID-19 has similar imaging features with other pneumonia viruses. Due to the lack of confirmation of each case, we do not add other viral pneumonia for comparison. Finally, CT images collected by different devices will hinder the promotion of the algorithm.

In our future work, we are going to increase annotated examples of COVID-19 and invite more experts to check the accuracy of the annotated data. Furthermore, we plan to add data on pneumonia caused by other types of viruses and retrain the data to improve the robustness of the algorithm. Finally, we will consider joining multiple institutions to collect chest CT images of various lesions of different severity for comparison to verify the effectiveness of the proposed method. We hope the proposed method can be promoted on a large scale clinically application and help diagnose other diseases.

Declaration of Competing Interest

We would like to declare that no conflict of interest exists in the submission of this manuscript, and this manuscript is approved by all authors for publication.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Nos. 62025603, U1705262, 61772443, 61572410, 61803234 and 61702136), Principal Fund (No. 20720200091) and Natural Science Foundation of Guangdong Province in China (No. 2019B1515120049).

References

[1] C. Wang, P.W. Horby, F.G. Hayden, G.F. Gao, A novel coronavirus outbreak of global health concern, Lancet 395 (10223) (2020) 470–473.
[2] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, et al., Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China, Lancet 395 (10223) (2020) 497–506.
[3] J. CSSE, Coronavirus COVID-19 global cases by the center for systems science and engineering (CSSE) at Johns Hopkins University (JHU), 2020.
[4] A. Berheim, X. Mei, M. Huang, Y. Yang, Z.A. Fayad, N. Zhang, K. Diao, B. Lin, X. Zhu, K. Li, et al., Chest CT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection, Radiology (2020) 200463.
[5] X. Xie, Z. Zhong, W. Zhao, C. Zheng, F. Wang, J. Wang, X. Liu, Chest CT for typical 2019-nCoV pneumonia: relationship to negative RT-PCR testing, Radiology (2020) 200343.
[6] A. Narin, C. Kaya, Z. Pamuk, Automatic detection of coronavirus disease (COVID-19) using x-ray images and deep convolutional neural networks, arXiv preprint arXiv:2003.10849 (2020).
[7] I.D. Apostolopoulos, T.A. Mpiesiana, COVID-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks, Phys. Eng. Sci. Med. (2020) 1.
[8] L. Wang, A. Wong, COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images, arXiv preprint arXiv:2003.09871 (2020).
[9] A. Oulefki, S. Agaian, T. Trongtrirakul, A.K. Laour, Automatic COVID-19 lung infected region segmentation and measurement using CT-scans images, Pattern Recognit. (2020) 107747.
[10] X. Chen, L. Yao, T. Zhou, J. Dong, Y. Zhang, Momentum contrastive learning for few-shot COVID-19 diagnosis from chest CT images, Pattern Recognit. (2020) 107626.
[11] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, L. Xia, Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases, Radiology (2020) 200642.
Using deep learning to detect pneumonia caused by nCoV-19 from x-ray images, available at: https://towardsdatascience.com/using-deeplearning-to-detect-nco-19-from-x-ray-images-a18970161aecd (2020).

Large-scale screening of COVID-19 from community acquired pneumonia using infection size-aware classification, Phys. Med. Biol. 66 (2021) 065031.

Lung infection quantification of COVID-19 in ct images with deep learning, arXiv preprint arXiv:2003.04653 (2020).

Fully convolutional networks for semantic segmentation, Proceedings of the IEEE International conference on computer vision and pattern recognition, 2015, pp. 3341–3340.

L.-C. Chen, G. Papandreou, F. Schroff, H. Adam, Rethinking atrous convolution for semantic image segmentation, arXiv preprint arXiv:1706.05587 (2019).

K. Hao, Z. Mao, X. Zhang, C. Guo, What can help pedestrian detection? in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3127–3136.

K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969.

N. Dvornik, K. Shenkelov, J. Mairal, C. Schmid, BrainNet: a real-time deep network for scene understanding, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4154–4162.

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

T. Song, Generative model-based ischemic stroke lesion segmentation, arXiv preprint arXiv:1906.02392 (2019).

D.-P. Fan, T. Zhou, G.-P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, L. Shao, MedNet: automatic COVID-19 lung infection segmentation from CT scans, arXiv preprint arXiv:2004.14133 (2020).

J.K. Bolden, T.S. Engdahl, S. Pedraza, T.-H. Cho, G. Thomalla, N. Nghoghossian, J.-C. Baron, J. Feihler, L. Østergaard, K. Mournier, Better diffusion segmentation in acute ischemic stroke through automatic tree learning anomaly segmentation, Front. Neuroinformatics 12 (2018) 21.

S.M. Abulnaga, J. Rubin, Ischemic stroke lesion segmentation in CT perfusion scans using pyramid pooling and focal loss, in: International MICCAI Brainlesion Workshop, Springer, 2018, pp. 352–363.

R. Zhang, L. Zhao, W. Lou, J.M. Abrigo, V.C. Mok, W.C. Chu, D. Wang, L. Shi, Automatic segmentation of acute ischemic stroke from DWI using 3-D fully convolutional DenseNet, Proc. SPIE, 2018, pp. 732–734.

L. Liu, L. Kurgan, F.-X. Wu, J. Wang, Attention convolutional neural network for accurate segmentation and quantification of lesions in ischemic stroke disease, Med. Image Anal. 65 (2020) 101791.

S. Wang, D. Wu, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.

W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: single shot multibox detector, in: European Conference on Computer Vision, Springer, 2016, pp. 21–37.

C. Fu, W. Liu, A. Ranga, A. Tyagi, A.C. Berg, DSSD: deep convolutional single shot detector, arXiv preprint arXiv:1701.06659 (2017).
Xi Li received the B.M/Bachelor of Medicine degree in Wuhan Medicine from Wuhan University, Hubei, Wuhan, China in 2001 and the M.D (medical doctor) degree from Wuhan University, Hubei in 2006. He is now working in the Shenzhen Hospital of Peking University, Shenzhen, China. His current research interests include artificial intelligence research about medical pathology and CT imaging.

Jie Chen received the MSc and Ph.D degrees from the Harbin Institute of Technology, China, in 2002 and 2007, respectively. He joined the faculty with the Graduate School in Shenzhen, Peking University, in 2019, where he is currently an associate professor with the School of Electronic and Computer Engineering. Since 2018, he has been working with the Peng Cheng Laboratory, China. From 2007 to 2018, he worked as a senior researcher with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. In 2012 and 2015, he visited the Computer Vision Laboratory, University of Maryland and School of Electrical and Computer Engineering, Duke University respectively. His research interests include pattern recognition, computer vision, machine learning, deep learning, and medical image analysis. He is an associate editor of the Visual Computer. He is a member of the IEEE.

Jiawen Zheng obtained a B.Sc. in School of Information Science & Engineering, Lanzhou University, Lanzhou, China. He is currently studying for a M.Sc. at the School of Informatics Xiamen University, Xiamen, China. His main research includes deep learning, computer vision and pose estimation.

Yongsheng Liang received the Ph.D. degree in communication and information systems from the Harbin Institute of Technology, Harbin, China, in 1999. From 2002 to 2005, he was a Research Assistant with the Harbin Institute of Technology, where he has been a Professor with the School of Electronic and Information Engineering, since 2017. His research interests include video coding and transferring and machine learning and applications. He was the recipient of the Wu Wenjun Artificial Intelligence Science and Technology Award for Excellence and the Second Prize in Science and Technology of Guangdong Province.

Yonghong Tian (Senior Member, IEEE) is currently a Full Professor with the School of Electronics Engineering and Computer Science, Peking University, Beijing, China, and the Deputy Director of AI Research Center, Pengcheng Lab, Shenzhen, China. His research interests include computer vision, multimedia big data, and brain-inspired computation. He is the author or coauthor of more than 170 technical articles in refereed journals and conferences. He was/is an Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (2018), IEEE TRANSACTIONS ON MULTIMEDIA (2014/2018), IEEE Multimedia Magazine (2018), and IEEE Access (2017). He co-initiated the IEEE International Conference on Multimedia Big Data (BigMM) and served as the TPC Co-Chair of BigMM 2015, and also served as the Technical Program Co-Chair of IEEE ICME 2015, IEEE ISM 2015 and IEEE MIPR 2018/2019, and the organizing committee member of ACM Multimedia 2009, IEEE MMSP 2011, IEEE ISCAS 2013, IEEE ISM 2015/2016. He is the steering member of IEEE ICME (2018) and IEEE BigMM (2015), and is a TPC Member of more than ten conferences such as CVPR, ICCV, ACM KDD, AAAI, ACM MM and ECCV. He was the recipient of the Chinese National Science Foundation for Distinguished Young Scholars in 2018, two National Science and Technology Awards and three ministerial-level awards in China, and received the 2015 EURASIP Best Paper Award for the EURASIP Journal on Image and Video Processing and the Best Paper Award of IEEE BigMM 2018. He is a senior member of CIE and CCF and a member of ACM.

Hui Sun received the B.Eng degree from Harbin Institute of Technology, Harbin, China, in 2005, and the M.E. degree from Tsinghua University, Beijing, China, in 2008. He is currently working toward the Ph.D. degree with the Department of Automation, Tsinghua University, Beijing, China. His research interests include multimedia information retrieval, machine learning, and pattern recognition.

Rongrong Ji (Senior Member, IEEE) is currently a Professor and the Director of the Intelligent Multimedia Technology Laboratory, and the Dean Assistant with the School of Information Science and Engineering, Xiamen University, Xiamen, China. His work mainly focuses on innovative technologies for multimedia signal processing, computer vision, and pattern recognition, with over 100 papers published in international journals and conferences. He is a member of the ACM. He also serves as a program committee member for several Tier-1 international conferences. He was a recipient of the ACM Multimedia Best Paper Award and the Best Thesis Award of Harbin Institute of Technology. He serves as an Associate/Guest Editor for international journals and magazines, such as Neurocomputing, Signal Processing, Multimedia Tools and Applications, the IEEE Multimedia Magazine, and Multimedia Systems.