2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

A Denoising Self-supervised Approach for COVID-19 Pneumonia Lesion Segmentation with Limited Annotated CT Images

Yibo Gao, Huan Wang, Xinglong Liu, Ning Huang, Guotai Wang and Shaoting Zhang

***Abstract*— The coronavirus disease 2019 (COVID-19) has become a global pandemic. The segmentation of COVID-19 pneumonia lesions from CT images is important in quantitative evaluation and assessment of the infection. Though many deep learning segmentation methods have been proposed, the performance is limited when pixel-level annotations are hard to obtain. In order to alleviate the performance limitation brought by the lack of pixel-level annotation in COVID-19 pneumonia lesion segmentation task, we construct a denoising self-supervised framework, which is composed of a pretext denoising task and a downstream segmentation task. Through the pretext denoising task, the semantic features from massive unlabelled data are learned in an unsupervised manner, so as to provide additional supervisory signal for the downstream segmentation task. Experimental results showed that our method can effectively leverage unlabelled images to improve the segmentation performance, and outperformed reconstruction-based self-supervised learning when only a small set of training images are annotated.**

***Clinical relevance*—The proposed method can effectively leverage unlabelled images to improve the performance for COVID-19 pneumonia lesion segmentation when only a small set of CT images are annotated.**

I. INTRODUCTION

The coronavirus disease 2019 (COVID-19) was first recognized in December 2019. Due to its contagious nature and lack of appropriate vaccines, it has been rapidly spreading to most countries worldwide and developed into a global pandemic [1]. As is reported by the center for systems science and engineering at Johns Hopkins University, there are globally 150 million cases confirmed COVID-19, with 3 million global deaths (updated 1 May, 2021) [2].

COVID-19 is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [3], whose common symptoms include cough, fever and shortness of breath and pneumonia [4]. Though reverse transcription polymerase chain reaction (RT-PCR) has become one of the golden standards in terms of COVID-19 diagnosis, it is shown to have a high false negative rate due to the practical issues in sample collection and transportation.

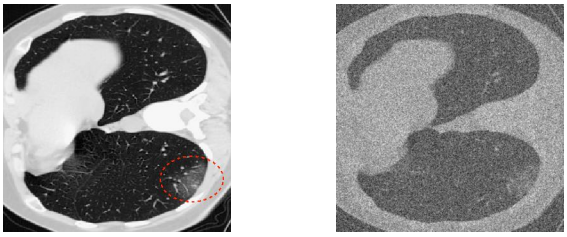
Medical images can be used as a complementary tool for detecting and evaluating COVID-19 infections [5], [6].

This work was supported by Glasgow College, University of Electronic Science and Technology of China and the National Natural Science Foundation of China under Grant 81771921 and Grant 61901084. Yibo Gao and Huan Wang contributed equally to this work. Corresponding author: Guotai Wang (guotai.wang@uestc.edu.cn).

Yibo Gao, Huan Wang and Guotai Wang are with the University of Electronic Science and Technology of China, Chengdu, China.

Xinglong Liu and Ning Huang are with SenseTime Research, Shanghai, China.

Shaoting Zhao is with the University of Electronic Science and Technology of China and SenseTime Research.



(a) (b)

Fig. 1. (a) Clean CT image of COVID-19 patient. (b) CT image of COVID-19 patient with Gaussian noise. The pneumonia GGO lesion is indicated in the red circle.

The typical pneumonia lesions such as Ground-Glass Opacity (GGO) in the early stage and pulmonary consolidation in the late stage [7] could be observed from Computed Tomography (CT) slices, as shown in **Fig. 1.(a)**. What's more, the segmentation of pneumonia lesions is important in the assessment of COVID-19 patients. Nevertheless, the manual segmentation for 3D volumes is laborious and time-consuming. In addition, manual segmentation of CT images is also a subjective task, since it can be easily influenced by the clinical experience and individual bias of the radiologists. Thus, an effective automatic segmentation method is highly required for COVID-19 in clinical practice.

After the outbreak of COVID-19, a lot of methods based on deep learning have been proposed for segmenting the CT images of COVID-19 [8], [9], [10]. Despite of the powerful ability of deep neural networks to learn visual features, their performance greatly depends on the scale of training data. Since obtaining sufficient pixel-level annotated data for segmentation is quite expensive and infeasible during the pandemic time, the performance of those data-driven methods is limited.

In order to alleviate the shortage of annotated medical images, this paper aims to propose a feature learning algorithm in a self-supervised manner. Self-supervised mechanism can learn features from unannotated data effectively by exploiting the internal structure of data. The core of the self-supervised framework is to design an appropriate pretext to fully exploit massive unlabelled data. [11]. Since most networks follow the encoder-decoder design paradigm for medical image segmentation, the architecture can also be used in image denoising task [12]. In our work, we find that forcing the network to reduce the Gaussian noise added on the COVID-19 CT images facilitates the internal feature learning. Thus, we propose a self-supervised framework for COVID-19 lesion segmentation by image denoising.

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

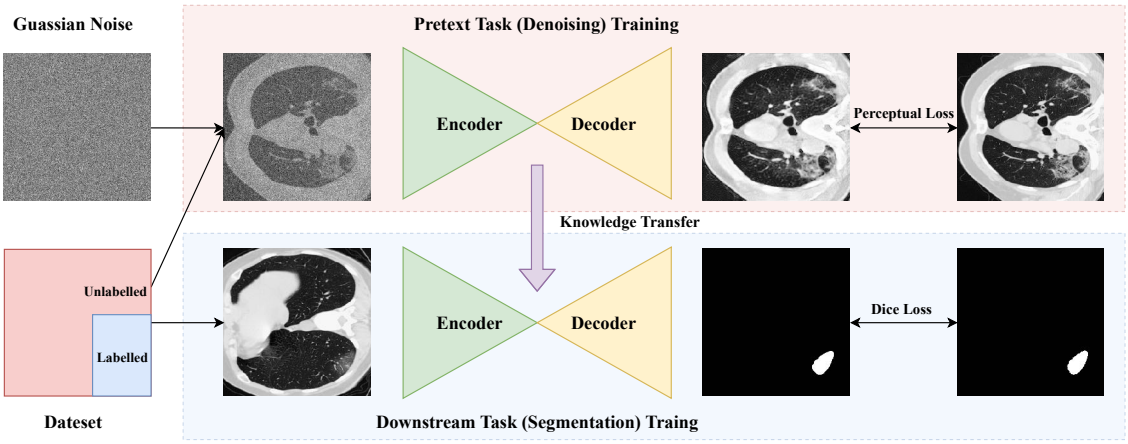


Fig. 2. The pipeline of the proposed self-supervised framework.

The framework proposes to perform the denoising pretext task, where the semantic features from massive unlabelled data can be extracted. Then, the features are transferred to the downstream task to enhance the segmentation performance when only limited annotations are available. Our underlying hypothesis is that doing well on the denoising task requires networks to learn the internal visual features of CT images. If networks do well in the denoising task, they must extract the critical features of the pneumonia lesion from massive unlabelled data, which will provide surrogate supervisory signal for the segmentation task. In this way, the performance limitation due to the lack of annotated data can be alleviated. Most of the states-of-the-arts image segmentation networks utilize an encoder-decoder architecture, which can also be used in image denoising tasks. Therefore, the proposed framework can be implemented with different network structures.

In this paper, we explore the way to utilize massive unlabelled data and construct a framework tailored for COVID-19 lesion segmentation in a self-supervised method, so as to alleviate the shortage of annotated data. We evaluate our method by extensive experiments for different amount of annotated data and different segmentation backbone networks. For the most frequently used UNet architecture, on a testing set of 50 cases, with 220 cases for denoising pretext task and 40 cases for downstream segmentation task, the framework improves the volume-level dice by 1.25%, which shows the effectiveness of our method.

II. Method

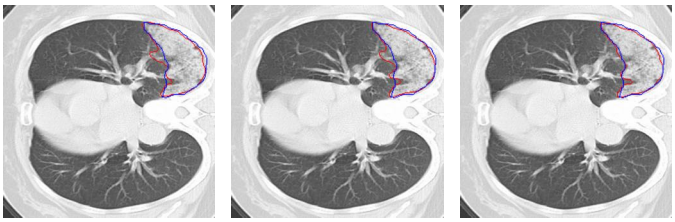
For COVID-19 segmentation task, due to laborious manual segmentation, only a small set of images are annotated, which means the given dataset D can be divided into the labelled set $D_l = \{(X_1, Y_1), (X_2, Y_2), ..., (X_l, Y_l)\}$ and the unlabelled set $D_u = \{X_{l+1}, X_{l+2}, ..., X_{l+u}\}$, where $l \ll u$. Thus, we explore the method to fully utilize unlabelled data for segmenting COVID-19 lesion and construct a pipeline for the task. As shown in **Fig. 2**, the pipeline consists of two tasks which are the pretext denoising task and the downstream segmentation task. In the pretext task, we add Gaussian noise to unlabelled data for denoising. Then, the noised and original images are used as training data and labels respectively to train the network with perceptual loss [13]. In order to transfer the features learned from the massive unlabelled data, we fine-tune the network using annotated images in a supervised method. With the learned features during the pre-training for the pretext task, the performance of the downstream segmentation task can be greatly improved. The training procedure of the framework is summarized in **Algorithm 1**.

Algorithm 1 The Training Procedure of the Framework

Input: Network with scratch parameters $F(\theta)$,
labelled dataset $D_l = \{(X_1, Y_1), (X_2, Y_2), ..., (X_l, Y_l)\}$,
unlabelled dataset $D_u = \{X_{l+1}, X_{l+2}, ..., X_{l+u}\}$
Output: Optimized network $F(\theta^*)$
for each $X_i \in D_u$ **do**
  add Gaussian noise to image $X_i \leftarrow N(X_i|\boldsymbol{\mu}, \Sigma^2)$;
  calculate the output $F(X_i|\theta)$;
  calculate the perceptual loss L_p as Eq(2);
  update the parameter of the network $\theta \leftarrow \theta - \alpha \frac{\partial L_p}{\partial \theta}$
end
for each $(X_i, Y_i) \in D_l$ **do**
  calculate the output $F(X_i|\theta)$;
  calculate the dice loss $L_d = 1 - \frac{2 \cdot Y_i \cdot F(X_i|\theta) + \epsilon}{Y_i + F(X_i|\theta) + \epsilon}$;
  update the parameter of the network $\theta \leftarrow \theta - \alpha \frac{\partial L_d}{\partial \theta}$
end

A. Pretext Denoising Task

The objective of the pretext task is to learn the semantic features in an unsupervised paradigm. In order to achieve this objective, we propose to train the segmentation network $F(\cdot)$ to reduce the artifacts applied to the images $X \in D$. Specifically, we denote $N(\cdot|\boldsymbol{\mu}, \Sigma^2)$ to be the operator to apply Gaussian noise, where $\boldsymbol{\mu}$ and Σ^2 are the mean vector and covariance matrix of the Gaussian distribution. The segmentation network $F(\cdot)$ gets the noised image $X' = N(X|\boldsymbol{\mu}, \Sigma^2)$ as input and yields the denoised image as output $I = F(X'|\theta)$, where θ is the parameters of network



(a) Fully-supervised (b) Self-reconstruction (c) Self-denoising

Fig. 3. Visual comparison of segmentation results obtained by different methods when only 40 volumes in the training set were annotated.Red and blue curves indicate the network predictions and the labelled mask, respectively.

$F(\cdot)$. Therefore, the training objective function of the pretext task is:

$$\min_{\theta} \frac{1}{u+l} \sum_{i=1}^{u+l} L_d(I_i, X_i). \tag{1}$$

Typically, Mean Squared Error (MSE) loss function is frequently used in the denoising task, trying to minimize the pixel-wise error between the network output and the original image. However, MSE can bring about blurs and distortions in output images, which hinders the network to learn semantic features from the pretext task. Therefore, we employ a perceptual loss [13] to make the network learn more high-level features, so as to gain more details from the noised image. Besides, a total variance regularization is also added to smoothen the output image. The total loss function for the denoising task is given below:

$$\begin{aligned} L_d = & \frac{1}{w_F h_F d_F} \left\| VGG(I) - VGG(X) \right\|_1 \\ & + \lambda \frac{1}{w_I h_I} \sum_{i,j} ((I_{i,j-1} - I_{i,j})^2 + (I_{i+1,j} - I_{i,j})^2). \end{aligned} \tag{2}$$

Here $VGG()$ is a feature extractor implemented by the pre-trained VGG-19 [14] on ImageNet, where we obtain the feature map from the 16th layer of VGG-19. It is worth noting that we pad the 3 input channels of VGG-19 by the original single channel CT image for implementation. w_F, h_F and d_F represent the width, height and depth of the obtained feature map. λ is to control the relative weights between perceptual loss and total variance regularization. w_I and h_I are the width and height of the network output I.

B. Downstream Segmentation Task

In the downstream task training phase, we follow the supervised learning paradigm and train the network with a small amount of labelled data D_l to make the features learned in the pretext task applicable for COVID-19 segmentation. For the sample (X_i, Y_i) from D_l, X_i and Y_i represent the original CT slice and the segmentation mask respectively. In the pretext training phase, we extract critical features from massive unlabelled data with the network $F(\cdot|\theta)$. In order to transfer the features we obtained and facilitate the segmentation, we use (X_i, Y_i) to update the parameters in the network by minimizing the dice loss:

$$\min_{\theta} \frac{1}{l} \sum_{i=u+1}^{u+l} DiceLoss(F(X_i|\theta), Y_i) \tag{3}$$

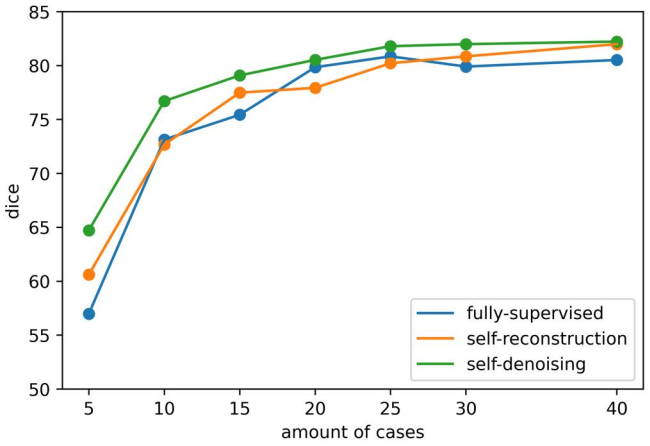


Fig. 4. Quantitative comparison of different methods with different numbers of annotated volumes for training.

III. Experiments and Results

A. Data and Implementation Settings

For the experiments, images of 330 COVID-19 patients, which have a large range of inter-slice spacing (0.625mm-8.0mm) and the pixel size (0.61mm-0.93mm). Since the images are of different sizes, we center crop and resize the images to the same size (256×256). The total 330 cases of the dataset were randomly split into 220, 40, 20 and 50 cases with 29772, 5330, 3230 and 5829 slices for unsupervised training, supervised training, validation and testing. Two experts manually annotated the ground truth masks with consensus. Since the amounts of slices for different cases are of great difference, the testing dice coefficient is calculated on a volume-level.

The experiments are conducted on a workstation with a NVIDIA RTX 2070 super. Adam optimizer is implemented with a mini-batch of 16, weight decay 10^{-5}, and cosine annealing. For both pretext and downstream training, the learning rate are set to 10^{-3} and 5×10^{-3}, the training epochs are 20 and 150 respectively.

B. Comparison for Different Amounts of Labelled Cases

In this experiment, we compare our denoising framework (self-denoising) with supervised method (fully-supervised) and the self-supervised method of image reconstruction (self-reconstruction). For the comparative method, the pretext task is to reconstruct the original images from images with a 64×64 randomly selected missing region. Both two self-supervised methods use the same experiment settings. With UNet [15] as the backbone segmentation network, we first implement the pretext task to learn from unannotated data. Then, the pre-trained model is fine-tuned to validate the effectiveness of the supervisory signal provided by the

TABLE I

DICE SCORES ACHIEVED BY DIFFERENT METHODS UNDER DIFFERENT NUMBERS OF ANNOTATED VOLUMES IN THE TRAINING SET

Method	5 cases	10 cases	15 cases	20 cases	25 cases	30 cases	40 cases
Fully-Supervised	56.98 ± 20.76	73.14 ± 15.34	75.44 ± 11.99	79.85 ± 11.07	80.86 ± 10.22	79.91 ± 10.56	80.52 ± 10.93
Self-Reconstruction	60.61 ± 20.13	72.68 ± 15.39	77.49 ± 12.98	77.94 ± 11.30	80.22 ± 10.96	80.86 ± 10.65	81.99 ± 10.12
Self-Denoising	$\mathbf{64.72 \pm 19.94}$	$\mathbf{76.71 \pm 12.95}$	$\mathbf{79.09 \pm 11.35}$	$\mathbf{80.53 \pm 11.94}$	$\mathbf{81.80 \pm 10.54}$	$\mathbf{81.99 \pm 9.97}$	$\mathbf{82.23 \pm 10.03}$

TABLE II

THE RESULTS OF DIFFERENT SEGMENTATION NETWORKS WITH 40 CASES IN SELF-DENOISING AND FULLY-SUPERVISED

Method	PSPNet	UNet	UNet++
Fully-Supervised	80.99 ± 10.97	80.52 ± 10.93	82.21 ± 10.21
Self-Denoising	$\mathbf{81.98 \pm 10.44}$	$\mathbf{82.23 \pm 10.03}$	$\mathbf{83.35 \pm 9.65}$

pretext task, using 5, 10, 15, 20, 25, 30 and 40 cases of annotated data respectively.

Fig. 3 demonstrates the qualitative comparison of the predictions generated by three different methods in the experiment. According to the results shown in **Table 1** and **Fig. 4**, we can observe that the proposed self-denoising framework can enhance the best segmentation results for all the amounts of cases, whereas the self-reconstruction approach fails to obtain better segmentations than fully-supervised method in three settings. This observation indicates emphasizes that the proposed pretext framework is appropriate and effective for COVID-19 segmentation task. When the number of annotated cases is 40, self-denoising get the best segmentation performance (Dice=82.23%). It is worth noting that as the amount of labelled data decreasing, the dice improvement increases, which demonstrates that the pretext task improves the network's feature representation ability by leveraging unlabeled images.

C. Effectiveness for Different Backbone Networks

Since our training method was not designed for a specific network structure, in this experiment, we verify the effectiveness of the framework for different backbone networks. Except for UNet, we additionally select two states-of-the-arts segmentation networks, which are PSPNet [16] and UNet++ [17]. All the labelled 40 cases containing 5330 slices are used in this experiment. It should be noted that all the experiment settings are the same except for the backbone network. The experimental results are presented in **Table 2.**

It can be seen that UNet++ obtains the best dice (83.35%) in both two methods, since PSPNet was initially proposed for natural image segmentation and UNet++ is an improved version of UNet. The proposed framework can improve all the three segmentation networks, so that our proposed self-supervised pipeline is general for different networks.

IV. CONCLUSION

In this work, we construct a self-supervised framework for COVID-19 segmentation task by image denoising. The pretext denoising task helps the encoder-decoder networks to utilize the abundant unlabelled data for the downstream segmentation task, so as to improve the segmentation performance. According to the experimental results, it can be observed that the framework can improve the segmentation performance in terms of different amount of labelled data and different segmentation networks.

REFERENCES

[1] Chen Wang, Peter W Horby, Frederick G Hayden, and George F Gao et al, "A Novel Coronavirus Outbreak of Global Health Concern," *The Lancet*, vol. 395, no. 10223, pp. 470 – 473, 2020.

[2] "Coronavirus COVID-19 Global Cases by the Center for Systems Science and Engineering at Johns Hopkins University," https://coronavirus.jhu.edu/map.html, Accessed: May 2, 2021.

[3] Chaolin Huang, Yeming Wang, Xingwang Li, Lili Ren, and Bin Cao, "Clinical Features of Patients Infected with 2019 Novel Coronavirus in Wuhan, China," *The Lancet*, vol. 395, no. 10223, pp. 497 – 506, 2020.

[4] Ming-Yen Ng, JElaine Y. P. Lee, in Yang, Fangfang Yang, Xia Li, Hongxia Wang, Macy Mei-sze Lui, Christine Shing-Yen Lo, Barry Leung, Pek-Lan Khong, Christopher Kim-Ming Hui, Kwok-yung Yuen and Michael D. Kuo, "Imaging Profile of the COVID-19 Infection: Radiologic Findings and Literature Review," *Radiology*, 2020.

[5] Zuhao Liu, Huan Wang, Wenhui Lei, and Guotai Wang, "CSAF-CNN: Cross-Layer Spatial Attention Map Fusion Network for Organ-at-Risk Segmentation in Head and Neck CT Images," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020, pp. 1522–1525.

[6] Zuhao Liu, Huan Wang, Shaoting Zhang, Guotai Wang, and Jin Qi, "NAS-SCAM: Neural Architecture Search-Based Spatial and Channel Joint Attention Module for Nuclei Semantic Segmentation and Classification," in *MICCAI*, 2020, pp. 263–272.

[7] Zheng Ye, Yun Zhang, Yi Wang, Zixiang Huang, and Bin Song et al, "Chest CT Manifestations of New Coronavirus Disease 2019 (COVID-19): a Pictorial Review," *European Radiology*, pp. 1–9, 2020.

[8] Guotai Wang, Xinglong Liu, Kang Li, and Shaoting Zhang et al, "A Noise-Robust Framework for Automatic Segmentation of COVID-19 Pneumonia Lesions From CT Images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2653–2663, 2020.

[9] Dengping Fan, Tao Zhou, Gepeng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen and Ling Shao, "Inf-net: Automatic Covid-19 Lung Infection Segmentation from CT Images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2626–2637, 2020.

[10] Titinunt Kitrungrotsakul, Qingqing Chen and Lanfen Lin et al, "Attention-RefNet: Interactive Attention Refinement Network for Infected Area Segmentation of COVID-19," *IEEE Journal of Biomedical and Health Informatics*, 2021.

[11] Longlong Jing and Yingli Tian, "Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[12] Zifei Yan, Shi Guo, Gang Xiao and Hongzhi Zhang, "On Combining CNN With Non-Local Self-Similarity Based Image Denoising Methods," *IEEE Access*, vol. 8, pp. 14789–14797, 2020.

[13] Justin Johnson, Alexandre Alahi and Feifei Li, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," in *ECCV*, 2016, pp. 694–711.

[14] Karen Simonyan and Andrew Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *ICLR*, 2015.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *MICCAI*, 2015, pp. 234–241.

[16] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid Scene Parsing Network," in *CVPR*, 2017.

[17] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang, "UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.