SC2Net: A Novel Segmentation-based Classification Network for Detection of COVID-19 in Chest X-ray Images

Zhenyu Fang †, Huimin Zhao †, Jinchang Ren, Calum MacLellan, Yong Xia, Shuo Li, Meijun Sun, and Kevin Ren

***Abstract*— The pandemic of COVID-19 has become a global crisis in public health, which has led to a massive number of deaths and severe economic degradation. To suppress the spread of COVID-19, accurate diagnosis at an early stage is crucial. As the popularly used real-time reverse transcriptase polymerase chain reaction (RT-PCR) swab test can be lengthy and inaccurate, chest screening with radiography imaging is still preferred. However, due to limited image data and the difficulty of the early-stage diagnosis, existing models suffer from ineffective feature extraction and poor network convergence and optimisation. To tackle these issues, a segmentation-based COVID-19 classification network, namely SC2Net, is proposed for effective detection of the COVID-19 from chest x-ray (CXR) images. The SC2Net consists of two subnets: a COVID-19 lung segmentation network (CLSeg), and a spatial attention network (SANet). In order to supress the interference from the background, the CLSeg is first applied to segment the lung region from the CXR. The segmented lung region is then fed to the SANet for classification and diagnosis of the COVID-19. As a shallow yet effective classifier, SANet takes the ResNet-18 as the feature extractor and enhances high-level feature via the proposed spatial attention module. For performance evaluation, the COVIDGR 1.0 dataset is used, which is a high-quality dataset with various severity levels of the COVID-19. Experimental results have shown that, our SC2Net has an average accuracy of 84.23% and an average F1 score of 81.31% in detection of COVID-19, outperforming several state-of-the-art approaches.**

***Index Terms*— COVID-19, chest x-ray imaging, SC2Ne, lung segmentation, ResNet-18**

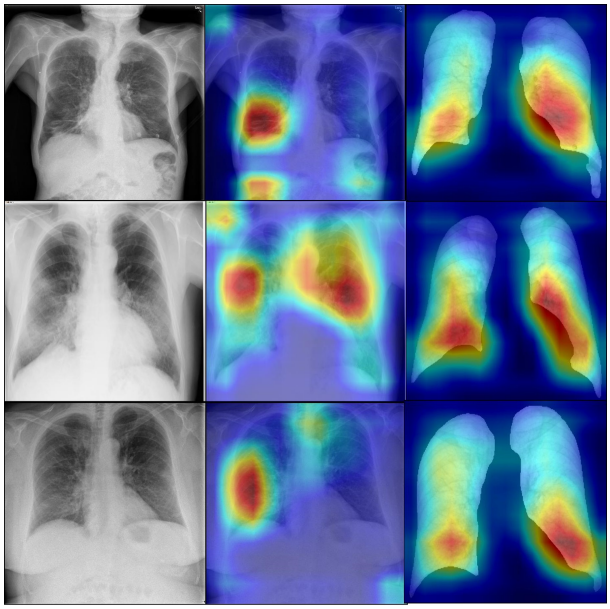


Fig. 1: A heat map shows the class-discriminating regions using GRAD-CAM++. Specifically, for the examples on the third columns, the size and the position are normalized with the background region suppressed. The severity levels from top to bottom rows are: MILD, MODERATE and SEVERE, respectively.

I. Introduction

SINCE January, 2020, the novel coronavirus pneumonia (COVID-19) has rapidly spread to more than 188 countries and regions, and caused more deaths, than the previous coronavirus strains, namely, Severe Acute Respiratory Syndrome (SARS) and the Middle East Respiratory Syndrome (MERS) [1]. As a result, many countries have taken substantial losses in terms of their population health [2]. It is therefore imperative in the fight against this disease for medical institutions to be equipped with the tools necessary for fast and effective diagnosis.

Currently, the gold standard for diagnosing of the COVID-19 is the real-time reverse transcriptase polymerase chain reaction (RT-PCR) swab test [3]. However, the diagnostic results from RT-PCR requires several hours to process, and

† The first two authors contribute equally to this work.
Z. Fang, H. Zhao and J. Ren are with the School of Computer Sciences, Guangdong Polytechnic Normal University, Guangzhou, China.
Z. Fang is also with the School of Computer Software, Northwestern Polytechnical University, Xi'an, China.
J. Ren, corresponding author, is with National Subsea Centre, Robert Gordon University, Aberdeen, UK, E-mail: jinchang.ren@ieee.org.
C. MacLellan is with the Dept. of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, UK.
Y. Xia is with the School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, China
Y. Li is with the School of Biomedical Engineering, Western University, London, ON, Canada
M. Sun is with the School of Computer Science and Technology, Tianjin University, Tianjin, China
K. Ren is with the School of Medicine and Dentistry, University of Aberdeen, Aberdeen, U.K.

studies have shown that the test suffers from a high false negative rate [4], often requiring repeated tests.

An alternative and quicker diagnostic method is based on chest radiography imaging (CRI), where patients with suspected COVID-19 symptoms may undergo computed tomography (CT) or chest X-ray (CXR) screening to visualize observable thoracic lesions. Some have argued that CT is more suitable for COVID-19 detection over CXR imaging [5]. However, CT imaging is expensive, time-consuming, and not always readily available. At the detriment of image resolution and contrast, CXR imaging presents a cheaper, quicker, and more easily accessible alternative to CT, with demonstrated high efficacy in detecting COVID-19 [6]. However, no matter the selected data modality, key radiological features such as ground-glass opacities, crazy-paving patterns, bilateral involvement, and peripheral distributions consistent with the COVID-19 are also partially presented in SARS and MERS [7]. With such a short time period for radiologists to build requested knowledge and experience, discriminating COVID-19 from other pneumonias becomes a challenging task. To this end, also taking the increased pressure on health services from rising cases into account, it is a need to develop robust computer-aided diagnosis systems for automating the diagnostic process and to ease the burden on clinical staff.

In recent years, deep convolutional neural networks (DCNN) have emerged as an effective tool for a wide range of image analysis tasks such as segmentation, classification, and object detection. DCNNs, or deep learning models, have actually learned the handcrafted features on their own by framing the learning process as an optimization problem. This has enabled them to iteratively determine the key class-discriminating distributions in the data without any manual intervene in prior. Having a model that can effectively leverage the data is therefore crucial to medical image analysis, which has motivated the wide deployment of deep learning in medical imaging and computer-aided diagnosis. For tacking the pandemic of the COVID-19, therefore, a deep learning-based computer-aided diagnosis (CAD) system can have the potential to learn from hundreds of imaging cases and thus provide an effective tool for improved decision-making in terms of COVID-19 detection and diagnosis. Recently, DCNN based methods have achieved promising performance on detecting COVID-19 [8], [9]. However, there still exists challenges on detect COVID-19 cases:

i. As seen in Figure 1, existing models detect the COVID-19 cases from the whole CXR directly, where the information from the background regions, e.g., arms and neck, may affect the training of the classifiers and lead to degraded accuracy and robustness, i.e., the models of COVID-19 may have the learning target altered due to background noise caused by non-lung regions (NLR) and variations of the size and position of the lung (SPL);
ii. As discussed in the previous works [10], [11], and also validated in our experiments, existing CNNs have difficulty on tackling data shift of CXR-based COVID-19 detection. In other words, the classifier may achieve a promising performance on one dataset, but fail when tested on another;
iii. Under limited training samples, to increase the depth of CNNs cannot necessarily improve the accuracy of COVID-19 diagnosis, thus a more effective model is needed to tackle this challenge;
iv. As illustrated in [10], existing methods focus mainly on the diagnosis of severe cases, where early-stage cases, which are hardly detectable, are seldom considered;

Motivated by this, we have proposed a segmentation based deep learning based COVID-19 network, namely SC2Net, to detect the COVID-19 from early-stage to late-stage, with a high reliability. A COVID-19 lung segmentation network is utilized to exclude the non-lung area and normalize the size of lung. Due to the limited number of samples, the size normalization may lead to overfitting. Thus, a bounding-box oscillation strategy is proposed, which slightly shift the position of lung during training. Meanwhile, a spatial attention module and a multi-scale learning strategy are proposed, to further enhance the efficacy of feature extraction.

The major contributions of this paper can be summarized as follows:

i). We propose a cascaded segmentation-classification network (SC2Net), to alleviate the interference of the background noise, which may cause learning alternation issue and vulnerability to data shift. This is achieved by utilizing a novel COVID-19 lung segmentation network (CLSeg) before classification, for suppressing the effect of the background region in the extracted features;
ii). A bounding-box oscillation strategy is proposed to the segmented lung region for normalizing the position of the segmented lung in order to avoid overfitting of the fixed lung position;
iii). When trained on small datasets, increasing the number of layers may bring no improvement on the classification accuracy. Thus, a shallow CNN, namely spatial attention network (SANet) is proposed, in which a novel spatial attention module (SAM) is utilized to enhance the discriminability of high-level features, followed by a multi-scale learning method for improved feature extraction;

The remaining parts of this paper are organized as follows. Section II briefly introduces the related work. The architecture and implementation detail of the proposed method are presented in Section III, followed by the experimental results discussed in Section IV. Finally, some concluding remarks are given in Section V. The abbreviations utilized in this paper are summarized in Table I.

II. Recent Work

Since the outbreak of the COVID-19, many deep learning-based methods and models have been proposed for its detection and diagnosis from medical images, especially the CT and CXR images. Initially, two-class solutions were focused whereby COVID-19 was distinguished from either healthy images [21], [22], or from lung infections with similar image features, such as Vir. pneumonia (viral pneumonia) [23] and others [24], [25]. Wang *et al.* [9] proposed one of the first deep learning models, COVID-Net, a CNN based model for discriminating CXR of COVID-19 patients from samples of

TABLE I: Abbreviations used in this paper.

Abbreviations	CXR	Pneum.	Vir. Pneum.	Bact. Pneum.	NLR	SPL	BBOX	Acc.	Spec.	Prec.	Rec.	Sens.	MACs	ROC	AUC
Description	chest X-ray image	pneumonia	viral pneumonia	bacterial pneumonia	non-lung region in CXR	size-position of lung	bounding box	accuracy	specificity	precision	recall	sensitivity	multiply–accumulate operations	receiver operating characteristic	area under curve

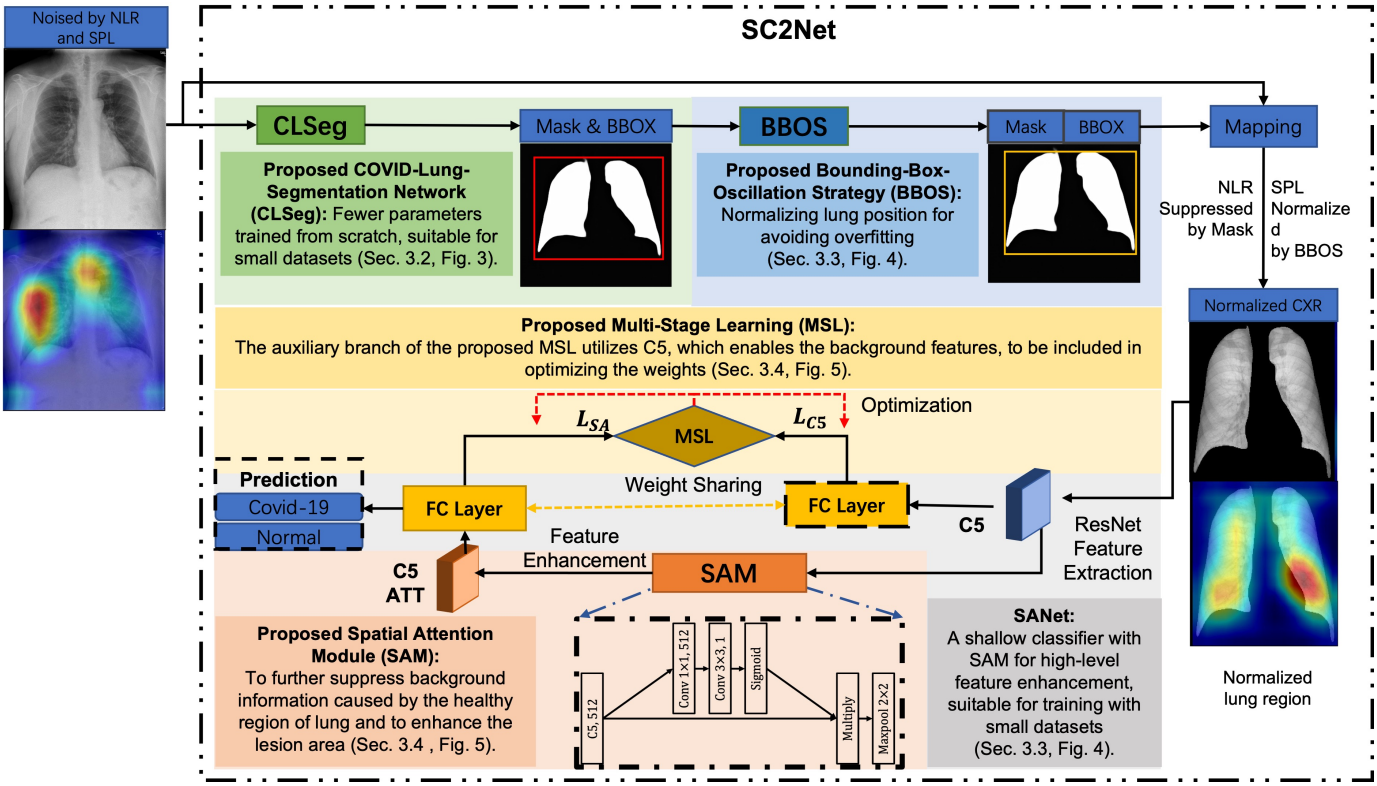


Fig. 2: Flowchart of the proposed SC2Net.

healthy, Vir. pneumonia, and Bact. pneumonia (bacterial pneumonia) patients. Using a dataset consisting of 1203 normal, 660 Vir. pneumonia, 931 Bact. pneumonia, and 45 COVID-19 patients, a testing accuracy of 83.5% was achieved. However, with such a large class imbalance, the reliability of COVID-Net for use on different datasets could hardly be assured.

Following this, Karim *et al.* [12] proposed the DeepCOVID-Explainer, which was built with an ensemble of two different classification networks, i.e., DenseNet-161 and VGG-19. By training their ensemble method on a dataset of 11,896 normal, pneumonia, and COVID-19 CXR images, they outperformed COVID-Net [9] with a precision and recall of 89.61% and 83.00% on a hold-out test set containing 77 COVID-19 samples. They also demonstrated the effectiveness of the Grad-CAM++ method [26] for highlighting the class-discriminating pixels on the test images, enabling more interpretable decisions of their models to the clinical doctors.

In [13], Bassi *et al.* attempted to overcome the class imbalance issue of limited COVID-19 images, where a transfer learning technique was used to leverage the pre-trained CheXNet model [8]. The model was trained on 127 COVID-19, 1285 pneumonia, and 1281 normal CXR images. They also implemented conventional data augmentation methods (e.g., flipping, translations, and rotations) to boost their sample sizes to 9144, 8128, and 8128 for the COVID-19, pneumonia, and normal cases, respectively. A tested mean accuracy of 97.8% was achieved.

Zhang *et al.* [14] proposed the so-called COVID-DA model, in which both labeled and unlabeled data were used to train the model in a semi-supervised fashion. In [15] another ensemble strategy was proposed to include three ResNets [27], where each of them was trained on a subset of their CXR training dataset for a separate binary classification. Once trained, the three models were stacked and fine-tuned on a fourth dataset. The multi-channel ensemble model could obtain a precision and recall values of 94% and 100% respectively, giving superior performance over an individual ResNet trained only on the fourth dataset.

A comparative study was conducted by Chawki *et al.* [16], where various network architectures used for deep learning based COVID-19 detection were implemented, including VGG-16, VGG-19, DenseNet-201, Inception-ResNet-V2, Inception-V3, ResNet-50, and MobileNet-V2 for classifying images as either normal, bacterial, viral, or COVID-19 pneumonia. Their dataset consisted of 1583 normal, 2780 bacterial, 1493 viral, and 231 COVID-19 cases. With an 80/20

TABLE II: Summary of methods and findings for approaches on Chest X-ray images.

Reference	Dataset	Method	Results
Karim [12]	COVID-19 259 Pneum. 8614 Normal 8066	DenseNet, ResNet, VGG19 (ensemble)	89.1% (Prec.) 83.0% (Rec.)
Bassi [13]	COVID-19 219 Vir. Pneum. 1345 Normal 1341	Pre-trained CheXNet (DenseNet-121)	98.3% (Acc.)
Zhang [14]	Training Testing COVID-19 258 60 Pneum. 2306 n/a Normal 8154 885	semi-supervised domain adaption. Uses a generator-discriminator design to learn pneumonia CXR data distributions to enable COVID-19 data distribution to be better distinguished.	0.9298 F1 score
Kim [15]	COVID-19 184 Pneum. 4245 Normal 1579	Ensemble three ResNets	94% (Prec.) 100% (Rec.)
Chawki [16]	COVID-19 231 Bact. Pneum. 2780 Vir. Pneum. 1493 Normal 1583	Inception-ResNetV2	92.18% (Acc)
Wang [9]	COVID-19 386 Pneum. 5551 Normal 8066	COVID-Net	95.9% (F1 score)
Tabik [10]	COVID-19 426 Normal 426	COVID-SDNet (GAN based)	81.00% (Acc.)
Tuncer [17]	COVID-19 135 Pneum. 150 Normal 150	a fuzzy transform (F-transform) is used as feature extractor, and 16 conventional classifiers are utilized.	97.01% (Acc.)
Hu [18]	Dataset 1: COVID-19. 520, non-COVID 5000 Dataset 2: COVID-19 219, Pneum. 4290, Normal. 1583	LetNet-5 is used as feature extractor. Extreme Learning Machines is applied as classifier.	98.25% (Acc. on dataset 1) 99.11% (Acc. on dataset 2)
Gilanie [19]	COVID-19 1066 Pneum. 7021 Normal 7021	CNN	96.68% (Acc.)
Afifi [20]	COVID-19 1056 Pneum. 5541 Control 7218	DenseNet-161	91.2% (Acc.)
Lin [11]	Dataset 1: COVID-19. 363, Pneum. 3736, Normal 1408 Dataset 2: COVID-19 617, Pneum. 5575, Normal. 8066 Dataset 3: COVID-19 427, Pneum. 426, Normal. 427	adaptive deformable ResNet	98.55% (Acc. on dataset 1) 95.00% (Acc. on dataset 2) 89.53% (Acc. on dataset 3)

training and testing split, the Inception-ResNet-V2 was found to outperform all other models, giving an overall accuracy of 92.18% and F1-score of 92.07%, respectively.

More recently, Tuncer et al. [17] proposed a lightweight multileveled feature extraction method via a fuzzy transform (F-transform) based on triangle fuzzy sets and a formed fuzzy tree. Taresh et al. [28] and Nayak et al. [29] utilized different pretrained CNNs for COVID-19 diagnosis. When classify COVID-19 cases from normal cases and viral pneumonia cases, VGG16 and MobileNet achieved the best performance [28]. ResNet-34 surpasses other CNNs, if viral pneumonia cases are not considered [29]. Hu et al. [18] replaced the fully connected layer in CNN by extreme learning machines. A new CNN architecture is proposed by Gilanie et al. [18], which can be applied to both chest X-Ray and CT images for diagnosis of COVID-19. For a more concise comparison of the literature, we have listed the discussed literature, together with their respective methods and results, in Table II.

Actually, attention modules have showed their particular strengthen and values on detection of COVID-19. In [30], a context attention network, formed by 3D depth-wise and 3D residual squeezing and excitation block, was proposed for segmenting the lesion region on CT scans. Sitaula et al. [31] utilized an attention module to enhance the spatial relationships between the regions of interests in CXR images. Afifi et al. [20] proposed a bi-path attention architecture, where both global and local deep features were adopted in a multi-label classification framework. Lin et al. [11] enhanced

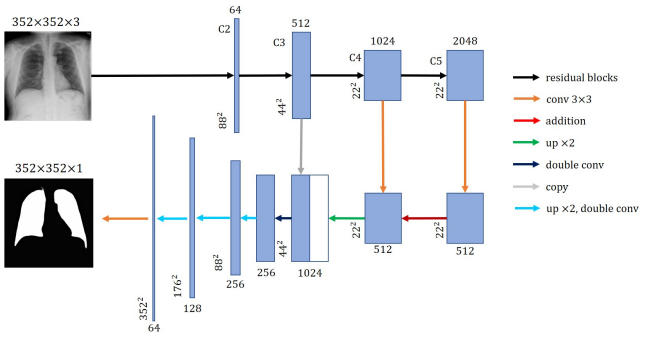
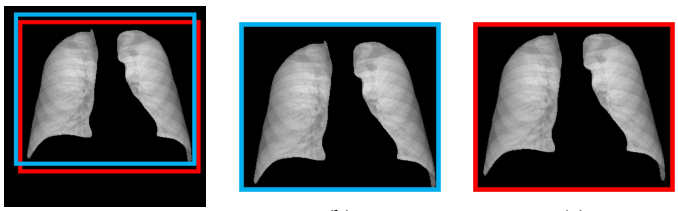


Fig. 3: Architecture of the proposed CLSeg model.



(a) (b) (c)

Fig. 4: The proposed bounding-box oscillation strategy (BBOS).

the efficacy of feature extraction via adaptive deformable convolution. Zhang et al. [32] proposed a multiple-input CNN, where the feature was enhanced by the convolutional block attention module. Zhou et al. [33] introduced two contrastive abnormal attention models to discover intra- and inter-contrastive abnormal between two lungs.

Although previous works have achieved promising performance in COVID-19 detection and diagnosis, experimental results in Maguolo *et al.* [34] showed that the background region may affect result prediction as well, causing the altered learning target of the network. In addition, Tabik *et al.* [10] argued that most of the previous methods only collected severe cases for training and testing, without considering the early-stage cases. To tackle these issues, in this paper, we aim to improve the feature representation capability of a shallow network, which will be more suitable for training on a small number of samples yet with high prediction reliability.

III. Proposed Method

In this section, we will discuss the proposed method in detail. At first, the overall classification pipeline will be introduced, followed by the architectures of the proposed segmentation network and the classification network.

A. Overall classification pipeline

The overall classification pipeline is shown in Figure 2, where ”NLR” and ”SPL” denote the noise caused by non-lung region and size-position of lung, respectively; ”BBOX” is the bounding-box containing the segmented lung region; ”CLSeg”, ”BBOS”, ”SAM” and ”SANet” denote the proposed COVID-19 lung segmentation network, bounding-box oscillation strategy, spatial attention module and the classifier enhanced by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3177854, IEEE Journal of Biomedical and Health Informatics

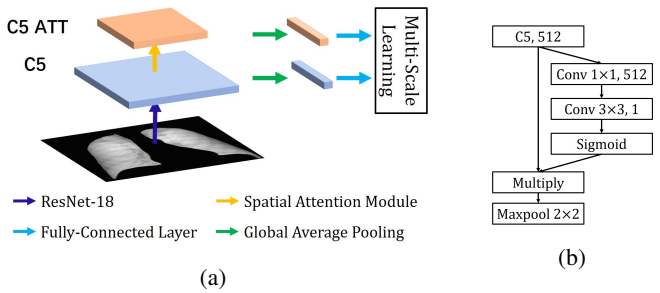


Fig. 5: Architecture of the proposed SANet (a) and spatial attention module (b), where the weights of the fully-connected layers are shared between two feature maps in the proposed multi-scale learning.

our spatial attention module, respectively; "C5" is the ResNet extracted features, and "C5 ATT" is the features enhanced by the proposed SAM. As visualized using Grad-CAM++, after background suppression, the classifier predicts based on the lung region only, without interfered by the background. Both C5 and C5 ATT are utilized for training, while in evaluation, only C5 ATT is utilized for prediction. The structure of the proposed CLSeg is detailed in Figure 3.

As illustrated in Tabik et.al [10], the CXRs produce images of the upper body, in which additional body parts may also be included, e.g., arms, neck, and stomach. The information from these additional body parts may inevitably affect the diagnosis of the COVID-19. To suppress the information of these background regions, the proposed segmentation network is first applied to segment the lung area from the plain CXR images, where the architecture will be detailed in Section III-B. Based on the predicted segmentation mask, a bounding box can be extracted to cover both the left and the right lung regions. The original CXRs are cropped using the generated bounding box, where non-lung area is excluded, and the sizes of the cropped CXRs are then unified for trained and evaluation. To further exclude the unwanted background information, the background pixels are filtered by multiplying the cropped image with the predicted segmentation mask. Cropping the CXR image by the extracted bounding box can help to normalize the position and the size of the lung region. However, the fixed position of the lungs constrains the variety of the dataset, especially when the number of samples is limited. As a result, the classifier can easily go overfitting during the training. To alleviate this position normalization caused problem, we proposed a bounding-box oscillation strategy (BBOS) on the cropping procedure, where the position of the lung is slightly shifted during cropping whilst remaining the whole lung region within the cropped image. The details of the proposed BBOS are presented in Section III-C. After surpassing the background information, CXR images are fed to the proposed SANet for classification and diagnosis of the COVID-19. The proposed SANet based classifier consists of a spatial attention module for feature enhancement, which will be detailed in Section III-D.

TABLE III: Result comparison (%) in terms of lung segmentation on the Cohen's dataset [35].

Methods	Dice	Precision	Sensitivity	Specificity
UNet	92.43	90.06	95.09	95.70
CLSeg	94.09	93.18	95.08	97.11

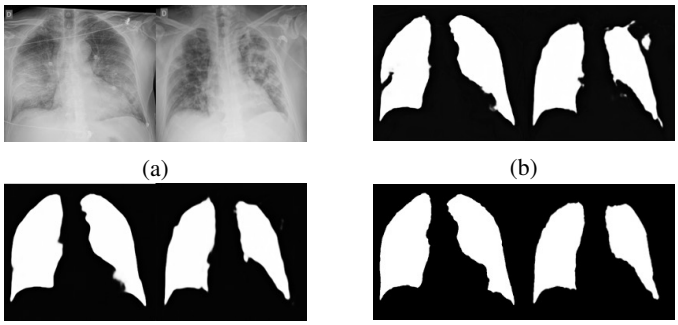


Fig. 6: Lung region segmentation results. The rows shown from top to bottom present: (a) original CXR images, (b) UNet prediction, (c) CLSeg prediction and (d) ground truth, respectively.

B. COVID-19 Lung segmentation network

The architecture of the proposed COVID-19 Lung segmentation network (CLSeg) is shown in Figure 3. Inspired by the U-Net [36], the proposed CLSeg consists of the encoder and decoder parts. The encoder of a plain U-Net is shallow, which constrains the capability of feature extraction. Thus, we adopt the pretrained ResNet, which are commonly used in medical imaging tasks [10], [37], as the encoder of the proposed CLSeg. As seen in Figure 3, the ResNet consists of 5 residual blocks, whose outputs are denoted respectively as "C1", "C2", "C3", "C4", and "C5". Specifically, the original ResNet reduces the feature map size by 32, i.e., the size of input image is 352×352, while the size of C5 is 11×11. According to experiments in FCN [38], such a small size is too coarse to capture local information of the high-level features. To tackle this drawback, we assign the downscale factor on the "layer 4" from 2 to 1. As a result, the downscale factor of the encoder becomes 16, which is identical to the U-Net. For the architecture of the decoder, the bilinear interpolation is utilized for upscaling the feature map. Similar to the U-Net, we adopt the feature fusion on the decoder. However, we did not directly concatenate the feature maps as used in the U-Net, simply because the channel size of the ResNet is much higher than the U-Net. As the decoder is trained from scratch, the increased parameters will impede the network optimization, especially for small datasets. Thus, the feature maps are fused via a mixture of addition and concatenation in the decoder part of the CLSeg. To further reduce the number of parameters, the channel size of C4 and C5 are normalized to 512 before addition. As suggested in the InfNet [37], low-level feature maps in the ResNet contribute little to the predicted result, hence the feature maps from C3 to C1 are excluded in the fusion. After applying the proposed CLSeg to the extracted

TABLE IV: Hyperparameters utilized in this paper.

	CLSeg	SANet
Image size	352	480
Optimizer	Adam	
Learning rate	1e-4	
Batch size	24	16
Epoch	100	10

lung region, the predicted map is produced as the outputted lung mask.

Let O and I respectively denote the prediction outputs and the CXR decision region of the classifier. For simplifying the discussions, only COVID-19 prediction task is considered in this section. Thus, the set of O is $\{+,-\}$, corresponding to the COVID-19 and normal cases, respectively. As the CXR-based COVID-19 diagnosis is achieved by determining the infected lung region, we assign the set of I as $\{l,u\}$, where l is the lung region and u is the remaining region of the upper body, respectively. Based on the assumptions above, the inference process of the previous works can be expressed by:

$$P\left(O \mid I\right) = P\left(O \mid l\right) P\left(l\right) + P\left(O \mid u\right) P\left(u\right) \tag{1}$$

After utilizing the CLSeg, the inference process becomes:

$$\begin{aligned} P\left(O \mid I\right) =& P\left(O \mid l \times F_{\text{CLseg}}\left(l\right)\right) P\left(l \times F_{\text{CLseg}}\left(l\right)\right) + \\ & P\left(O \mid u \times F_{\text{CLseg}}\left(u\right)\right) P\left(u \times F_{\text{CLseg}}\left(u\right)\right) \\ =& P\left(O \mid l\right) P\left(l\right) \end{aligned} \tag{2}$$

where F_{CLSeg} denotes the proposed CLSeg. The mask-based feature suppression can help to eliminate the prior of non-lung region. This enforces the decision region to be focused on the lung, resulting in more accurate and reliable diagnosis.

C. The proposed bounding-box oscillation strategy

With the CLSeg, the lung mask can be obtained from the image as the foreground, where the non-lung region will be considered as the background for diagnosing the COVID-19. As shown in Figure 2, the non-lung region may affect the learning of the classification model. Thus, the background information needs be suppressed. After background suppression, following the work in COVID-SDNet [10], the bounding-box is generated. The lung region is then cropped with the boundary of the cropped CXR image as the bounding-box, i.e., reassigning the lung region to the centre of the image. To further unify the size of the cropped lung region, the cropped CXRs are resized to 480×480 as suggested in [9]. However, with the normalized size and the position of the lung, the feature variety of each pixel is still limited, resulting in the overfitting of the classifier. This phenomenon is also observed in our experiments, which will be detailed in the next section.

To tackle the overfitting when normalizing the position and the scale of the lung region, we proposed a bounding-box oscillation strategy (BBOS). The bounding-box can be represented by (x, y, w, h), indicating its center coordinates, width and height, respectively. According to the definition of the convolution, a pixel with a coordinate $[m, n]$ on a feature map Y is determined by:

$$Y[m,n] = \sum_i \sum_j X[i,j] \cdot h[m-i, n-j] \tag{3}$$

where, X and h respectively denote the input and the convolution kernel. After utilizing the BBOS, the coordinate of the bounding box shifted by δ and becomes $(x \pm \delta, y \pm \delta, w, h)$. Meanwhile, the pixel $[m, n]$ on the feature map Y can be extracted by:

$$Y[m,n] = \sum_i \sum_j X[i \pm \delta, j \pm \delta] \cdot h[m-i, n-j] \tag{4}$$

An example is illustrated in Figure 4: the initial bounding-box,colored in red in (a), delimited the lung region, is acquired based on the segmentation result. In the proposed BBOS, the bounding-box is slightly shifted before cropping. Compared with the original center crop (see (c)), BBOS (see (b)) also maintains the whole lung region within the cropped image. However, the proposed BBOS remarkably increases the variety of samples.

D. Spatial attention module

After suppressing the background information, the CXR image is then fed to a classifier for COVID-19 diagnosis. As depicted in Section II, existing methods are still weak on diagnosing early-stage cases, i.e., at mild level. This is because the lesion area is still small in comparison to the healthy region. Thus, a spatial attention module (SAM) is proposed, to further suppress the background information caused by the healthy region of lung and to enhance the lesion area.

The architecture of the proposed SAM is shown in Figure 5. As seen, the proposed SAM is deployed to enhance the high-level feature map extracted via the pretrained ResNet-18. Here, a pretrained network is exploited due to the lack of sufficient training samples for the COVID-19 cases. Experimental results in [9], [10] found that, ImageNet pretrained networks outperforms the networks that are trained from scratch or via transfer learning. The process of feature enhancement in the SAM can be expressed by:

$$Y = \sigma\left(F_{3\times3}\left(F_{1\times1}(X)\right)\right) \times X \tag{5}$$

where X is the input feature map from layer 4; $F_{3\times3}$ and $F_{1\times1}$ are the convolution layers with the kernel sizes of 1×1 and 3×3, respectively; $\Sigma(*)$ is the sigmoid activation function and Y is the output of the proposed SAM.

After enhanced by the SAM, the values of the background pixels are further suppressed. However, those low-value pixels still affect the predicted result due to the used average pooling. Thus, the output of the SAM is further upsampled by 2, using a max-pooling layer. To improve the performance in terms of the fully connected layer, we propose a multi-scale learning in terms of the high-level feature maps as follows. During the training, the fully connected layer independently predicts the category using feature maps in both a high resolution (C5) and

TABLE V: Ablation study (%) of the proposed method on the COVIDGR 1.0 dataset.

Class	Metric	Methods				
		Baseline	+ Masked	+ BBOS	+ SA	+ MSL
Normal	Spec.	87.06 ± 7.10	$\mathbf{91.76 \pm 3.83}$	87.00 ± 7.83	86.47 ± 2.12	90.12 ± 4.45
	Prec.	78.91 ± 4.39	77.39 ± 4.56	82.63 ± 3.70	$\mathbf{83.53 \pm 3.76}$	83.07 ± 4.75
	F1	82.41 ± 1.41	83.78 ± 2.06	84.50 ± 3.99	84.88 ± 1.21	$\mathbf{86.29 \pm 2.75}$
COVID-19	Sens.	70.63 ± 1.05	66.52 ± 10.39	77.34 ± 6.70	$\mathbf{78.83 \pm 6.19}$	77.05 ± 8.17
	Prec.	83.34 ± 7.49	$\mathbf{87.42 \pm 3.88}$	84.03 ± 7.59	82.79 ± 1.46	86.89 ± 4.71
	F1	75.37 ± 4.43	74.85 ± 7.15	80.07 ± 3.89	80.60 ± 2.96	$\mathbf{81.31 \pm 4.80}$
Accuracy		79.64 ± 1.75	80.39 ± 3.42	82.65 ± 3.76	83.03 ± 1.88	$\mathbf{84.23 \pm 3.39}$

TABLE VI: Ablation study (%) of the proposed preprocessing method on the COVIDGR 1.0 dataset.

Class	Metric	Methods		
		w/o CLSeg	w/o Masking	Proposed
Normal	Spec.	81.18 ± 5.45	81.35 ± 9.92	$\mathbf{87.00 \pm 7.83}$
	Prec.	81.38 ± 4.24	$\mathbf{84.64 \pm 4.00}$	82.63 ± 3.70
	F1	81.01 ± 1.37	82.42 ± 4.26	$\mathbf{84.50 \pm 3.99}$
COVID-19	Sens.	76.70 ± 9.06	$\mathbf{81.21 \pm 7.95}$	77.34 ± 6.70
	Prec.	77.42 ± 3.74	79.52 ± 7.90	$\mathbf{84.03 \pm 7.59}$
	F1	76.73 ± 2.14	79.65 ± 3.21	$\mathbf{80.07 \pm 3.89}$
Accuracy		79.16 ± 1.25	81.29 ± 3.35	$\mathbf{82.65 \pm 3.76}$

a low resolution (SAM). We refer the classifier with the SAM module as SANet, and the loss of the SANet is defined by:

$$L = L_{SA} + L_{C5} \tag{6}$$

where L_{SA} and the L_{C5} denote the cross-entropy losses predicted using feature maps of the SAM and C5, respectively.

IV. Experimental Results

In this section, a series of ablation study will be conducted to validate the effectiveness of each module, followed by performance assessment of the proposed approach in comparison to a few state-of-the-art methods. In addition, the hyperparameters and the dataset will also be introduced.

A. Experimental setup and evaluation metrics

Classification related experiments are conducted on the COVIDGR 1.0 dataset [10] and COVIDx dataset [9], COVIDGR is a CXR dataset labeled by four highly trained radiologists from the Hospital Universitario Clínico San Cecilio, Granada, Spain. COVIDGR contains 852 images, equally distributed in two classes, i.e., COVID-19 and normal cases, respectively. According to the infection severity introduced in [41], COVID-19 CXR image is further annotated based on the RALE score [42] in four levels, i.e., Normal-PCR+, Mild, Moderate and Severe. Specifically, CXR image with positive PCR and annotated as "Normal" by expert radiologists are labeled with Normal-PCR+. Following the previous works [9], [10], we only consider the classification task of Normal vs COVID-19, due to the limited samples of each severity level.

To further validate the robustness of the proposed SC2Net, the results are also compared on the COVIDx dataset [9], which is another publicly available dataset with respect to the diagnosis of COVID-19 cases. The COVIDx dataset consists of 14003 images in three categories, i.e., 8066 normal, 5551 non-COVID19 infection (pneumonia) and 386 COVID-19, respectively, where the COVID-19 samples are collected from more than 266 COVID-19 patients. In total 300 images are randomly selected as the testing set, 100 per category, according to the patients' ID. For a particular patient, the associated data will be used either for training or testing, hence there is no overlapped data in this context. In addition, patient cases of normal and non-COVID19 pneumonia are collected from RSNA Pneumonia Detection Challenge dataset [9] The training and testing sets are randomly divided according to the patient ID, which means that for a particular patient, the associated data will be used either for training or testing, hence there is no overlapped data in this context. We train our models on the training dataset and evaluate the performance on the testing dataset.

As in Table IV, similar to the previous works [9], [10], the height and the width for training and evaluation are 480. Training images are randomly flipped with a probability of 0.5. The Adam [43] optimizer is adopted with an initial learning rate of 1e-4. The training lasts 10 epochs with a batch size of 16. The model with the highest accuracy on the validation set is selected for the testing and evaluation.

As discussed in [10], there exists a high degree of variations on dataset distributions between different training sets. For a fair evaluation, a five-fold cross validation is used in all the experiments. Each experiment uses 80% of COVIDGR-1.0 samples for training and the remaining 20% for testing, and the average results from five individual runs are used for comparison. In addition, 10% of each training set is utilized as the validation set.

The ResNet-18 is utilized as the baseline method, and all the experimental results are reported on the test set. We compare the results with several popularly used quantitative metrics, which include the sensitivity, specificity, precision, F1 score and Accuracy as defined below:

$$\begin{aligned}
\text{Sensitivity} &= \frac{TP}{TP+FN} \\
\text{Specificity} &= \frac{TN}{TN+FP} \\
\text{Precision} &= \frac{TP}{TP+FP} \\
F_1 \text{ Score} &= \frac{2 * \text{ Precision } * \text{ Sensitivity}}{\text{Precision } + \text{ Sensitivity}} \\
\text{Accuracy} &= \frac{TP+TN}{TP+FP+FN+TN}
\end{aligned} \tag{7}$$

where "TP" "TN", "FP" and "FN" denote the numbers of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3177854, IEEE Journal of Biomedical and Health Informatics

TABLE VII: Result comparison (%) to state-of-the-art methods on the COVIDGR 1.0 dataset.

Class	Metric	Methods					
		COVIDNet	COVID-CAPS	FuCiTNet	COVID-SDNet	MSRCovXNet	SC2Net
Normal	Spec.	83.42 ± 15.39	65.09 ± 10.51	82.63 ± 6.61	85.20 ± 5.38	82.35 ± 6.49	$\mathbf{90.12 \pm 4.45}$
	Prec.	69.73 ± 10.34	71.72 ± 5.57	79.94 ± 4.28	78.88 ± 3.89	85.12 ± 4.07	$\mathbf{83.07 \pm 4.75}$
	F1	74.45 ± 8.86	67.52 ± 5.29	81.05 ± 3.44	81.75 ± 2.74	83.46 ± 3.13	$\mathbf{86.29 \pm 2.75}$
COVID-19	Sens.	61.82 ± 22.44	73.31 ± 9.74	78.91 ± 5.88	76.80 ± 6.30	$\mathbf{82.01 \pm 5.76}$	77.05 ± 8.17
	Prec.	79.50 ± 11.47	68.40 ± 5.13	82.43 ± 5.43	84.23 ± 4.59	79.73 ± 5.04	$\mathbf{86.89 \pm 4.71}$
	F1	65.64 ± 15.90	70.20 ± 4.31	80.37 ± 3.16	80.07 ± 0.04	80.60 ± 2.83	$\mathbf{81.31 \pm 4.80}$
Accuracy		72.62 ± 7.6	69.20 ± 3.61	80.77 ± 3.15	81.00 ± 2.87	82.20 ± 2.83	$\mathbf{84.23 \pm 3.39}$

TABLE VIII: Results comparison in terms of F1 score (%) with several state-of-the-art deep learning models on the COVIDx test dataset.

Methods	F1 Score(%)		
	Normal	Pneumonia	COVID-19
ResNet-18* [27]	93.5	93.1	95.3
ResNet-50* [27]	93.3	93.9	95.9
Res2Net-50* [39]	93.7	94.9	**96.4**
ChexNet* [13]	**94.2**	94.9	95.9
COVID-Net [9]	92.5	91.6	95.9
MSRCovXNet [40]	94.2	95.4	**96.4**
SC2Net (proposed)	**94.7**	**95.4**	95.9

* Trained by author with 5 runs

samples as "True Positive", "True Negative", "False Positive" and "False Negative", respectively. Considering that the processing time is both device dependent and framework dependent, the number of multiply–accumulate operations (MACs) is reported for a fair comparison of the computational cost. In addition, the number of parameters is used to measure the sizes of corresponding models. As each model is evaluated in 25 runs, the averaged results and the standard deviation are used for comparison.

By using the re-cropped CXR image for training can significantly improve the training efficiency. Actually, the original CXR image is firstly pre-processed as introduced in the COVID-SDNet, i.e. centered cropping of the lung region with the size of the bounding-box extended by 2.5%. During the training, the cropped image is resized to 490×490 and then randomly cropped to 480×480, where the size of cropped CXR image is the same as previous works [9], [10]. As the bounding-box is extracted via the boundary of the lung region, directly shifting the bounding-box may exclude lesion region in the cropped CXR image and result in misled results of training. By expanding the original CXR image before the cropping, the randomly shifting of the lung region does not exclude the lung region outside the bounding-box. For performance evaluation, the bounding-box oscillation strategy (BBOS) is unnecessary. Thus, only the centered cropping is utilized at the last step.

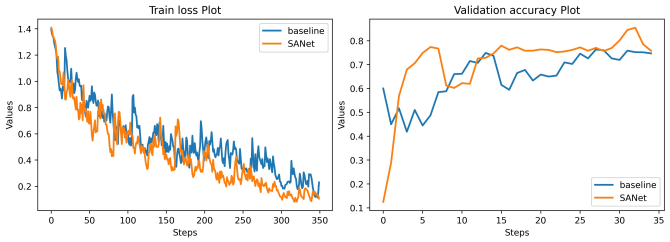
B. Comparison of the segmentation results

We train the proposed CLSeg on the dataset of Cohen et al. [35], using the hyperparameters as suggested in InfNet [37]. Apart from the precision, sensitivity, and specificity, the dice similarity coefficient (Dice), another golden metric in medical image segmentation [36], [37], is also used for comparison. The proposed CLSeg is compared with the standard UNet, which is used as the baseline method of COVIDGR dataset [10]. Experimental results on the dataset of Cohen et al. [35] are given in Table III, and the lung region segmentation results are shown in Figure 6. As seen, our CLSeg surpasses the UNet by 1.66% on the Dice. Specifically, the segmentation results from CLSeg are much closer than that of the UNet to the ground truth, as we have much less mis-segmented pixels. In contrast, results from the UNet are unsatisfactory, where a large number of lung pixels are mis-segmented. Additionally, the performance is reduced by 2% on the Dice, if feature maps are all fused via concatenation. This implies the efficacy of the mixed feature fusion method utilized in the CLSeg.

C. Effect of background suppression and the bounding-box oscillation strategy

The results of the ablation study are shown in Table V, where "+mask" implies the images are prepossessed by the lung mask, and "+BBOS" indicates the model is trained with our bounding-box oscillation strategy; "SA" and "MSL" denote the classifier using spatial attention module and the multi-scale leaning, respectively. Results of each column reported are from methods including its previous columns. As seen, the average accuracy is improved by about 3% after utilizing the background suppression and the proposed bounding-box oscillation strategy (BBOS). This has validated the effectiveness of the proposed pre-processing methods. Specifically, the average sensitivity of COVID-19 is improved by about 10% by utilizing the BBOS during training. This is because the BBOS enriches the varieties of local features in each layer, which alleviates the overfitting issue caused by the limited training samples. Furthermore, the gradient-guided class activation maps (Grad-CAM++) [26] is adopted to determine the class-discriminating regions in the input CXR image. As illustrated in Figure 1, the proposed background suppression method can focus the decision region on the lung, rather than other regions in the thoracic cavity. Thus, the proposed background suppression method could improve the reliability of the prediction, though it has only slightly increased the average accuracy by about 0.7%.

Effect of the proposed preprocessing method is shown in Table VI, where "w/o CLSeg" implies the images are masked and cropped by the lung mask generated using UNet; "w/o Masking" indicates the images are trained with cropping only,



(a) (b)

Fig. 7: Plots of the training loss (a) and the validation accuracy (b) of the proposed SANet, respectively.

of which the mask is generated via CLSeg. The proposed method utilizes masking and cropping as in the third column of Table V. We notice that by adding the mask for background suppression, the average sensitivity is improved by 4.1%. This is because, as illustrated in [34], when being evaluated on the original CXRs, the prediction may be affected by the background region, causing unreliable predictions. This has also been validated in Figure 1. However, by applying the masking, background information has been remarkably suppressed, leading to much improved average precision and more reliable prediction of the results. When utilized crop without masking, as seen in Table VI, the average accuracy drops by 1.4%, which further validates the necessary of the proposed BBOX for masking.

As discussed above, masking is essential in the proposed method. Accurate segmentation is then vial important for COVID-19 diagnosis. To validate the effect of the proposed CLSeg on classification of COVID-19, we adopt the masks predicted by UNet, with all others remaining the same. As shown in Table VI, the average accuracy of CLSeg surpasses the method using UNet by 3.5%. This is caused by the ineffective background suppression of UNet: as shown in Table III, the precision and low specificity of UNet is lower than CLSeg, which indicates that more false positive pixels are generated by UNet. As a result, the prediction using the UNet-masked CXR images suffers more interference by background region, than using the CLSeg-masked CXR images.

D. Improvements over the SANet

As seen in Table V, by utilizing the spatial attention (SA) module and the multi-scale learning, the average accuracies are further improved by 0.4% and 1.2%, respectively. As suggested in [44], [45], the auxiliary learning brings extra benefits to model training, where, the feature sizes of the main prediction branch and the auxiliary branch are the same. We have also trained the model with the identical feature size on those two branches. However, the advantages of the equal feature size in the two branches are not validated in our experiments. After the max pooling layer, the non-local maximum features will be discarded in forward propagation, which are also not utilized in back propagation. Thus, the compressed features will not contribute to the training. For comparison, the auxiliary branch of the proposed MSL utilizes C5, which enables the background features to be included in optimizing the weights.

TABLE IX: Performance (%) of the SC2Net with different classifiers.

Class	Metric	Methods		
		ResNet-18	ResNet-50	DenseNet-121
Normal	Spec.	87.00 ± 7.83	86.47 ± 8.07	$\mathbf{87.47 \pm 6.41}$
	Prec.	82.63 ± 3.70	$\mathbf{82.42 \pm 3.64}$	82.22 ± 3.25
	F1	84.50 ± 3.99	84.06 ± 3.67	$\mathbf{84.52 \pm 2.36}$
COVID-19	Sens.	$\mathbf{77.34 \pm 6.70}$	76.99 ± 6.88	76.48 ± 5.94
	Prec.	84.03 ± 7.59	83.38 ± 6.95	$\mathbf{84.09 \pm 5.58}$
	F1	$\mathbf{80.07 \pm 3.89}$	79.58 ± 3.36	79.75 ± 2.19
Accuracy		$\mathbf{82.65 \pm 3.76}$	82.20 ± 3.23	82.52 ± 2.02

TABLE X: Result comparison of computational cost.

Methods	Metric	
	#Params (M)	MACs (G)
ResNet-18	11.18	8.35
ResNet-50	23.51	18.87
COVIDNet	500.84	22.84
FuCiTNet*	11.95	195.45
COVID-SDNet*	41.56	357.67
MSRCovXNet	20.43	10.43
ClSeg	46.60	157.36
SANet	11.18	8.35
SC2Net*	57.78	165.71

* indicates the preprocessing module is applied

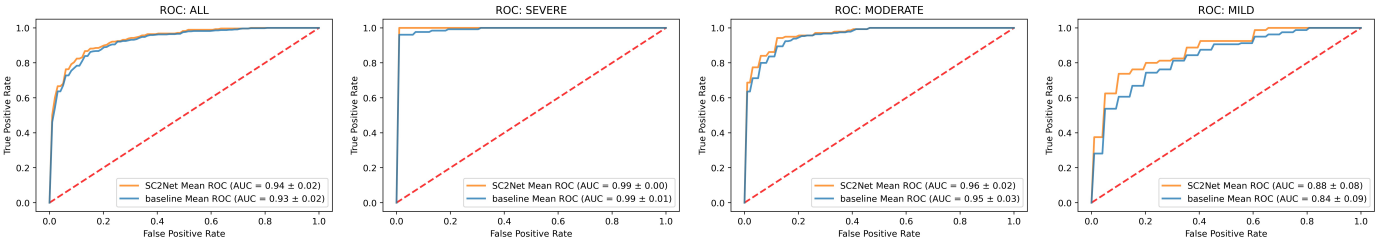
As seen in Figure 7, by utilizing the proposed spatial attention module and multi-scale learning, the training loss drops further by a margin, i.e., indicated reduced training difficulty. Meanwhile, the validation accuracy increases by at least 5%. This has clearly validated the effectiveness of the proposed SANet on small dataset whilst overfitting is avoided.

E. Effect on different severity levels

The ROCs of the proposed SC2Net, in comparison with the baseline method, are shown in Figure 8. When considering all three severity levels, the superiority of the proposed SC2Net over the baseline is insignificant, due mainly to the low difficulty in classifying severe cases. To validate this, we further plot the ROC for each severity level. As the severity level goes light, the gap between the proposed model and the baseline becomes obvious. For the mild level, the proposed SC2Net surpass the baseline by 4% on AUC, which indicates the effectiveness of the proposed method on diagnosis of COVID-19 in early-stage.

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

In this subsection, we compare our proposed SC2Net in terms of COVID-19 detection and diagnosis with several state-of-the-art methods, and the results are compared in Table VII. As seen, the proposed SC2Net outperforms the existing methods by at least 3% in terms of the average accuracy. When compared to the FuCiTNet and the COVID-SDNet, the precision, on the COVID-19 class, of our SC2Net surpasses those methods by 4.8% on average. This further validates the effectiveness of the background suppression components utilized in the SC2Net. It is worth noting that the COVID-SDNet uses the ResNet-50 as the backbone, while the FuCiTNet and SC2Net have a quite narrower backbone, the ResNet-18. To



(a) (b) (c) (d)

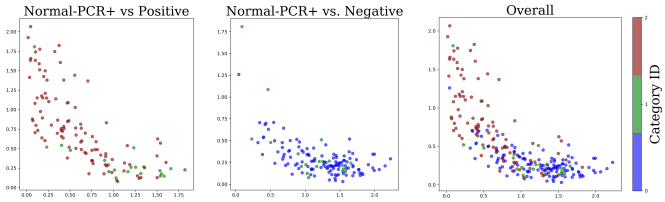
Fig. 8: ROC plots of the proposed SC2Net on the COVIDGR dataset: (a) all three severity levels, (b) severe, (c) moderate, (d) mild, where only the COVID-19 (positive), though in three severity levels, and the normal (negative) classes are used.

TABLE XI: Cross-dataset evaluation on the COVIDx test dataset.

Methods		F1 Score(%)	
		Normal	COVID-19
ResNet-18	plain	37.3	68.4
	+CLSeg	77.9	80.0
COVIDNet	plain	70.2	60.5
	+CLSeg	72.6	77.8
MSRCovXNet	plain	46.9	69.8
	+CLSeg	56.8	70.6
SC2Net	-	**80.8**	**80.2**

TABLE XII: Result comparison (%) to the models trained with or without Normal-PCR+, where Normal-PCR+ samples are included in both evaluations

Class	Metric	Methods	
		w/o. Normal-PCR+	w. Normal-PCR+
Normal	Spec.	**83.47 ± 10.56**	82.41 ± 0.88
	Prec.	**75.78 ± 4.33**	73.05 ± 4.40
	F1	**78.88 ± 4.56**	77.03 ± 3.69
COVID-19	Sens.	**72.45 ± 8.83**	68.73 ± 9.20
	Prec.	**82.86 ± 7.84**	80.71 ± 6.40
	F1	**76.51 ± 4.11**	73.49 ± 4.48
Accuracy		**77.97 ± 3.53**	75.58 ± 3.16



(a) (b) (c)

Fig. 9: Effect of Normal-PCR+ samples on classification.

explore further how the depth of the classifiers may affect the predicted results, we replace the backbone of the SANet with the three commonly used classifiers, i.e., ResNet-18, ResNet-50, and DenseNet-121. According to the results given in Table IX, the deeper network does not necessarily improve the classification accuracy. In contrast, the overall performances of the three classifiers are quite similar. Therefore, we deduce that the performance of deeper networks is constrained by the limited varieties of the samples, which could be improved when more high-quality samples are collected.

We compared the proposed SC2Net with other state-of-the-art deep learning models on the COVIDx dataset. The MSRCovXNet, also proposed by us, is compared with other methods that are trained on the same COVIDx dataset. Experimental results are shown and compared in Table VIII. The proposed SC2Net has achieved the state-of-the-art performance in classification of the Normal and Pneumonia classes, and comparable performance on COVID-19 class, which has validates its efficacy. Although the F1 score on the COVID-19 class is 0.5% lower than that of MSRCovXNet, we deduce that this is caused by the highly biased distribution of the severity levels. As illustrated in [10], the main severity level of COVID-19 cases in the COVIDx dataset is “Severe”, which is unsuitable for demonstrating the efficacy of the proposed method in detecting early stage cases at a lower severity level. As seen in TableVII, with the balanced data distribution, our SC2Net outperforms MSRCovXNet by 2% on accuracy, which further validates its robustness on detection of early stage COVID-19 cases.

Furthermore, we compared the MACs and the number of model parameters in Table X. In comparison with the FuCiTNet and COVID-SDNet that adopt the preprocessing modules before classification, SC2Net can reduce the computational cost by 15.22% and 53.67%, respectively. When compared with methods with only the classification module, i.e., ResNet, COVIDNet and MSRCovXNet, the proposed SC2Net needs much more MACs. However, about 94.96% of the overall computational cost is spent on CLSeg, which is essential for producing high-confidence prediction. If excluding the costs for segmentation, the proposed SANet only consumes as much as ResNet-18 in data classification. As a result, the cost of the proposed classifier module, i.e., SANet, is much lower than most of the methods without any preprocessing module.

G. Effect on data shift

In addition, we conduct cross-dataset evaluation to further validate the robustness of the proposed method. Following the same procedure as suggested in [10], we train our model on the COVID-GR dataset yet test it on the COVIDx dataset. This is because, in comparison to the COVIDx dataset, the COVID-GR dataset collects wider severity levels, though in fewer samples. For a fair comparison, we only report results on the COVD-19 and normal categories, as shown in Table XI.

By training on a different dataset, the F1 scores of the recently published methods all seem to degrade, due to the data shift caused by the settings during the CXR collection procedure [9], [10], [34]. However, the proposed SC2Net, thanks for the proposed spatial attention and background suppression modules, can still achieve state-of-the-art performance, outperforming existing works by at least 2% on detecting the COVID-19.

We have also applied the proposed CLSeg on the ResNet-18 and COVIDNet models, where the F1 scores in both cases are increased by a large margin. We deduce that the interference of background noise may be magnified on cross-dataset evaluation, as the data shift can affect both the foreground and the background regions. By combining with the proposed segmentation module, i.e., CLSeg, the background noise can be effectively suppressed for alleviating the interference caused by data shift hence the much-improved detection accuracy even for existing models such as ResNet-18 and COVIDNet. This further validate the efficacy of the proposed CLSeg on tackling the data shift issue, as well as the robustness of the proposed SC2Net.

H. Impact of Normal-PCR+ samples

As pointed out in the COVID-SDNet [10], the accuracy in diagnosing the Mild and Moderate severity levels is degraded if the Normal-PCR+ is absent. We also train the proposed method with Normal-PCR+ samples, where the result is reported in Table XII. The average accuracy is found to decrease by 2.4% when the Normal-PCR+ samples are used for training. The image features, extracted by the last hidden layer of the ResNet-18, are visualized in Figure 9. Values of the top node which contributes most to each category are selected for feature visualization. To be more specific, each sample is expressed by a pair of coordinates (x, y), where x and y are the values of node which contribute the most to the normal class and COVID-19 class, respectively. As x or y grows, the classier tends to output "Normal" or "COVID-19". The category IDs, for 0 and 1, correspond to the severity levels of Normal and Normal-PCR+, respectively. The category ID 2 contains samples with Mild, Moderate and Severe cases. Samples are selected from the validation set and the test set of the COVIDGR dataset (first fold, first partition). As seen, the feature differentiation between the Normal and the last two levels, Moderate and Severe, are significant. The features of the Mild level lie between the normal and the last two levels, which can be considered as hard samples. For majority of the Normal-PCR+ samples, however, they are mixed with the Normal samples. As illustrated in the COVIDGR dataset [10], there are no visual differences between the Normal and the Normal-PCR+ samples. However, we deduce that, by taking the Normal-PCR+ samples, as extremely hard samples into training, may be imped the weight optimization.

I. Limitations and future directions

The proposed SC2Net can be developed into a robust computer-aided diagnosis system for automating the diagnostic process and helping to ease the burden on clinical staff. To further provide the more detailed information when the severity accumulates, we will work on the prediction of the severity of the COVID-19 as well as the analysis of the dynamic pathology of the lesion in the lung regions.

At present, the segmentation module is utilized to alleviate the learning target alternation which has been misled by the background region. By further segmenting the lesion region, rather than whole lung region, the efficacy can be more improved. Meanwhile, the segmentation module and the classification module of the proposed SC2Net are in cascade, resulting in high computational burden. To tackle these limitations, the future work will focus on the prediction of the lesion region and parallel implementation of the segmentation and classification modules. Although the data shift issue is alleviated by the proposed CLSeg, the detection accuracy still lags the single-dataset-trained method with a large margin. As a result, transformation between datasets will also be explored in the future.

V. CONCLUSION

Detecting the COVID-19 at the early stage is essentially important for reducing the damage on patients and alleviating the burden on the clinical staff. At present, building a relatively large dataset of high clinical quality, consisting of equally distributed severity levels, is still a very challenging task. Accordingly, it is necessary to build a COVID-19 detector under limited samples, which is the major motivation of our proposed cascaded segmentation-classification method, SC2Net. Different from the previous works, the proposed SC2Net takes all severity levels into consideration, especially for early sage of COVID-19 samples.

As the lesion region is small at the early stage of COVID-19 CXRs and suffers from the interference of the background noise, a novel COVID-19 lung segmentation network, namely CLSeg, is utilized to effectively suppress the background of CXR. Different from the UNet, the proposed CLSeg module contains fewer training parameters and sparser encoder-coder connections, resulting in improved efficacy in dealing with small datasets. To alleviate the overfitting problem under limited training samples, a bounding-box oscillation strategy (BBOS) is proposed, by augmenting the local feature of the training sample. Moreover, a spatial attention module (SAM), in conjunction with a multi-scale learning method, is proposed to enhance the foreground feature on the high-level feature map. The architecture of the proposed SAM is lightweight, enabling high detection efficiency and efficacy. By embedding on a shallow backbone, the proposed SANet can outperform deep CNNs under a small number of training samples. Experimental results on COVIDx and COVIDGR datasets have validated the efficacy and robustness of the proposed approach in effective detection of the COVID-19 from CXR, there are still rooms for further improvements, such as working on the prediction of the severity of the COVID-19 and the analysis of the dynamic pathology of the lesion in the lung regions.

ACKNOWLEDGMENT

This work was supported in part by the Dazhi Scholarship of the Guangdong Polytechnic Normal University, the Key Lab-

oratory of the Education Department of Guangdong Province (2019KSYS009), the National Natural Science Foundation of China (62072122, 62006049, 61876125).

REFERENCES

[1] M. Elisabeth, "Coronavirus: covid-19 has killed more people than sars and mers combined, despite lower case fatality rate," https://www.bmj.com/content/368/bmj.m641.

[2] L. Pan, M. Mu, P. Yang *et al.*, "Clinical Characteristics of COVID-19 Patients With Digestive Symptoms in Hubei, China: A Descriptive, Cross-Sectional, Multicenter Study."

[3] M. M. Candace and Daniel, *COVID-19*, 2020. [Online]. Available: https://radiopaedia.org/articles/covid-19-3?lang=us

[4] J. F.-W. Chan, S. Yuan, K.-H. Kok *et al.*, "A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster," *The Lancet*, vol. 395, no. 10223, pp. 514–523, 2020.

[5] X. Qian *et al.*, "M3 lung-sys: A deep learning system for multi-class lung pneumonia screening from ct imaging," *IEEE journal of biomedical and health informatics*, vol. 24, no. 12, pp. 3539–3550, 2020.

[6] F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, K. He, Y. Shi, and D. Shen, "Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for COVID-19," *IEEE Reviews in Biomedical Engineering*, 2020.

[7] M. Chung, A. Bernheim, X. Mei, N. Zhang, M. Huang, X. Zeng, J. Cui *et al.*, "CT imaging features of 2019 novel coronavirus (2019-nCoV)," *Radiology*, vol. 295, no. 1, pp. 202–207, 2020.

[8] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya *et al.*, "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," *arXiv preprint arXiv:1711.05225*, 2017.

[9] L. Wang and A. Wong, "COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images," *arXiv preprint arXiv:2003.09871*, 2020.

[10] S. Tabik, A. Gómez-Ríos, J. L. Martín-Rodríguez *et al.*, "COVIDGR Dataset and COVID-SDNet Methodology for Predicting COVID-19 Based on Chest X-Ray Images," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 12, pp. 3595–3605, 2020.

[11] Z. Lin *et al.*, "Aanet: Adaptive attention network for covid-19 detection from chest x-ray images," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4781–4792, 2021.

[12] M. Karim, T. Döhmen *et al.*, "DeepCOVIDExplainer: Explainable COVID-19 Predictions Based on Chest X-ray Images," *arXiv preprint arXiv:2004.04582*, 2020.

[13] P. R. Bassi and R. Attux, "A Deep Convolutional Neural Network for COVID-19 Detection Using Chest X-Rays," *arXiv preprint arXiv:2005.01578*, 2020.

[14] Y. Zhang *et al.*, "COVID-DA: Deep Domain Adaptation from Typical Pneumonia to COVID-19," *arXiv preprint arXiv:2005.01577*, 2020.

[15] S. Misra, S. Jeon, S. Lee, R. Managuli, and C. Kim, "Multi-Channel Transfer Learning of Chest X-ray Images for Screening of COVID-19," *arXiv preprint arXiv:2005.05576*, 2020.

[16] K. Elasnaoui and Y. Chawki, "Using X-ray images and deep learning for automated detection of coronavirus disease," *Journal of Biomolecular Structure and Dynamics*, vol. 0, no. 0, pp. 1–12, 2020.

[17] T. Tuncer, F. Ozyurt, S. Dogan, and A. Subasi, "A novel covid-19 and pneumonia classification method based on f-transform," *Chemometrics and Intelligent Laboratory Systems*, vol. 210, p. 104256, 2021.

[18] T. Hu *et al.*, "Real-time covid-19 diagnosis from x-ray images using deep cnn and extreme learning machines stabilized by chimp optimization algorithm," *Biomedical Signal Processing and Control*, vol. 68, p. 102764, 2021.

[19] G. Gilanie *et al.*, "Coronavirus (covid-19) detection from chest radiology images using convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 66, p. 102490, 2021.

[20] A. Afifi *et al.*, "An ensemble of global and local-attention based convolutional neural networks for covid-19 diagnosis on chest x-ray images," *Symmetry*, vol. 13, no. 1, p. 113, 2021.

[21] O. Gozes, M. Frid-Adar *et al.*, "Rapid AI Development Cycle for the Coronavirus (COVID-19) Pandemic: Initial Results for Automated Detection & Patient Monitoring using Deep Learning CT Image Analysis," *arXiv preprint arXiv:2003.05037*, 2020.

[22] S. Asif and Y. Wenhui, "Automatic Detection of COVID-19 Using X-ray Images with Deep Convolutional Neural Networks and Machine Learning," *medRxiv*, 2020.

[23] S. Wang, B. Kang, J. Ma *et al.*, "A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19)," *European radiology*, pp. 1–9, 2021.

[24] B. Ghoshal and A. Tucker, "Estimating Uncertainty and Interpretability in Deep Learning for Coronavirus (COVID-19) Detection," *arXiv preprint arXiv:2003.10769*, 2020.

[25] B. Wang, S. Jin, Q. Yan *et al.*, "AI-assisted CT imaging analysis for COVID-19 screening: Building and deploying a medical AI system," *Applied Soft Computing*, vol. 98, p. 106897, 2021.

[26] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 839–847.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[28] M. M. Taresh, N. Zhu, T. A. A. Ali, A. S. Hameed, and M. L. Mutar, "Transfer learning to detect covid-19 automatically from x-ray images using convolutional neural networks," *International Journal of Biomedical Imaging*, vol. 2021, 2021.

[29] S. R. Nayak, D. R. Nayak, U. Sinha, V. Arora, and R. B. Pachori, "Application of deep learning techniques for detection of covid-19 cases using chest x-ray images: A comprehensive study," *Biomedical Signal Processing and Control*, vol. 64, p. 102365, 2021.

[30] A. Qayyum *et al.*, "Multilevel depth-wise context attention network with atrous mechanism for segmentation of covid19 affected regions," *Neural Computing and Applications*, pp. 1–13, 2021.

[31] C. Sitaula and M. B. Hossain, "Attention-based vgg-16 model for covid-19 chest x-ray image classification," *Applied Intelligence*, vol. 51, no. 5, pp. 2850–2863, 2021.

[32] Y.-D. Zhang *et al.*, "Midcan: A multiple input deep convolutional attention network for covid-19 diagnosis based on chest ct and chest x-ray," *Pattern recognition letters*, vol. 150, pp. 8–16, 2021.

[33] Y. Zhou *et al.*, "Contrast-attentive thoracic disease recognition with dual-weighting graph reasoning," *IEEE Transactions on Medical Imaging*, vol. 40, no. 4, pp. 1196–1206, 2021.

[34] G. Maguolo and L. Nanni, "A Critic Evaluation of Methods for COVID-19 Automatic Detection from X-Ray Images," *arXiv preprint arXiv:2004.12823*, 2020.

[35] J. P. Cohen *et al.*, "COVID-19 Image Data Collection," *arXiv 2003.11597*, 2020. [Online]. Available: https://github.com/ieee8023/covid-chestxray-dataset

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

[37] D.-P. Fan, T. Zhou, G.-P. Ji *et al.*, "Inf-Net: Automatic COVID-19 Lung Infection Segmentation From CT Images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2626–2637, 2020.

[38] J. Long *et al.*, "Fully Convolutional Networks for Semantic Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[39] S. Gao *et al.*, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[40] Z. Fang, J. Ren *et al.*, "A Novel Multi-stage Residual Feature Fusion Network for Detection of COVID-19 in Chest X-ray Images," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, pp. 1–1, 2021.

[41] H. Y. F. Wong, H. Y. S. Lam *et al.*, "Frequency and Distribution of Chest Radiographic Findings in Patients Positive for COVID-19," *Radiology*, vol. 296, no. 2, pp. E72–E78, 2020.

[42] M. A. Warren, Z. Zhao, T. Koyama *et al.*, "Severity scoring of lung oedema on the chest radiograph is associated with clinical outcomes in ARDS," *Thorax*, vol. 73, no. 9, pp. 840–846, 2018.

[43] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[44] Z. Fang, J. Ren *et al.*, "Triple loss for hard face detection," *Neurocomputing*, vol. 398, pp. 20–30, 2020.

[45] H. Liu *et al.*, "DARTS: differentiable architecture search," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.