SCOAT-Net: A Novel Network for Segmenting COVID-19 Lung Opacification from CT Images

Shixuan Zhao, Zhidan Li, Yang Chen, Wei Zhao, Xingzhi Xie, Jun Liu*, Di Zhao*, and Yongjie Li*

***Abstract*— The new coronavirus disease 2019 (COVID19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has spread worldwide at a rapid rate. There is no clinically automated tool to quantify the infection of COVID-19 patients. Automatic segmentation of lung opacification from computed tomography (CT) images provides excellent potential, which is of great significance for judging the disease development and treatment response of the patients. However, the segmentation of lung opacification from CT slices still faces some challenges, including the complexity and variability features of the opacity regions, the small difference between the infected and healthy tissues, and the noise of CT images. Besides, due to the limited medical resources, it is impractical to obtain a large amount of data in a short time, which further hinders the training of deep learning models. To answer these challenges, we proposed a novel spatial and channel-wise coarse-to-fine attention network (SCOAT-Net) inspired by the biological vision mechanism, which is for the segmentation of COVID-19 lung opacification from CT Images. SCOAT-Net has the spatial-wise attention module and the channel-wise attention module to attract the self-attention learning of the network, which serves to extract the practical features at the pixel and channel level successfully. Experiments show that our proposed SCOAT-Net achieves better results compared to state-of-the-art image segmentation networks.**

***Index Terms*— COVID-19, convolution neural network, segmentation, lung opacification, attention mechanism**

I. Introduction

THE new coronavirus disease 2019 (COVID-19) caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has become a continuing pandemic [1]–[4]. As of 9 September 2020, there have been 212 countries with outbreaks, a total of 27,486,960 cases diagnosed, and 894,983 deaths, even though the number of infected people continues to increase [5]. Clinically, reverse transcription-polymerase chain reaction (PT-PCR) is the gold standard for diagnosing COVID-19 [6], but it also has the disadvantages of a high false-negative rate [7]–[9] and the inability to understand the patient's condition.

COVID-19 has certain typical visible imaging features, such as lung opacification caused by ground-glass opacities (GGO), consolidation and pulmonary fibrosis, which can be presented in thoracic computed tomography (CT) images [9]–[11]. Therefore, CT can be used as an essential tool for clinical diagnosis. At the same time, CT can also directly reflect the changes in lung inflammation during the treatment process and is a crucial indicator for evaluating the treatment effect [6]. However, in the course of treatment, repeated inspections lead to a sharp increase in the workload of radiologists. Also, the assessment of inflammation requires a comparison of the region of lesions before and after treatment. Quantitative diagnosis by radiologists is inefficient and subjective and is difficult to be widely promoted. Artificial intelligence (AI) technology gradually plays an important role, it can perform fast and accurately becoming a powerful weapon for medical personnel. It can also realize the rapid response in multi-aspect such as diagnosis [12], segmentation [13], and quantitative analysis [14], assisting doctors in rapid screening, differential diagnosis, disease course tracking, and efficacy evaluation to improve the ability to handle the COVID-19. In this research, we focus on segmentation of COVID-19 lung opacification from CT images.

Benefit from the rapid development of deep learning [15], many excellent convolution neural networks (CNNs) have been applied to medical image analysis tasks and have achieved the most advanced performance [12], [16], [17]. CNN can be applied in various image segmentation tasks due to its excellent expression ability and a data-driven adaptive feature extraction model. However, its success is inseparable from the expert's accurate labeling of a large number of training images, making it not suitable for all tasks. COVID-19 lung opacification segmentation based on CT images is still an arduous task, which has the following problems: 1) In the emergency of the COVID-19 outbreak, it is difficult to obtain enough data with

This work was supported by Key Area R&D Program of Guangdong province with Grant No.2018B03033801, National Key Research and Development Project (2018ZX10723203-001-002), Key Emergency Project of Pneumonia Epidemic of novel coronavirus infection (2020SK3006), Emergency Project of Prevention and Control for COVID-19 of Central South University (160260005) and Foundation from Changsha Scientific and Technical bureau, China (kq2001001).

*Corresponding author: Jun Liu (junliu123@csu.edu.cn), Di Zhao (zhaodi@ict.ac.cn), and Yongjie Li (liyj@uestc.edu.cn).

Shixuan Zhao, Zhidan Li, and Yongjie Li are with the MOE Key Lab for Neuroinformation, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China.

Yang Chen is with the West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, China.

Wei Zhao, Xingzhi Xie and Jun Liu is with the Department of Radiology, The Second Xiangya Hospital, Central South University, No.139 Middle Renmin Road, Changsha, Hunan, China

Jun Liu is with the Department of Radiology Quality Control Center, Changsha, Hunan, China

Di Zhao is with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

accurate labels to train the deep learning models in a short time due to the limited medical resources. 2) The infection areas in the CT slice show various features such as different sizes, positions, and textures, and there is no distinct boundary, which increases the difficulty of segmentation. 3) Due to the complexity of the medical images, the lung opacity area is quite similar to other lung tissues and structures, making it challenging to identify. Severed works [18]–[20] have tried to solve these challenges from the perspectives of reducing manual depiction time, using noisy labels and semi-supervised learning, and achieved specific results. Our research is derived from the attention learning mechanism, which makes full use of the inherent extraordinary self-attention ability of CNN to make the network generate attention maps and the attention vectors in the training process to weight the spatial domain feature and channel domain feature respectively. The areas and features activated by the network can diagnose the target area more accurately. There have been a series of studies [21]–[23] that proved the effectiveness of the attention mechanism for the classification and segmentation tasks.

Attention mechanism stems from the study of biological vision mechanism [24], and selective attention is one of the characteristics of human vision. In cognitive neuroscience, it is believed that an individual cannot receive and pay attention to all stimuli due to the bottleneck of information processing. Humans will selectively focus on some information while ignoring other visible information. The feature integration theory proposed by Treisman [25] uses a spotlight to describe the spatial selectivity of attention metaphorically. This model points out that visual processing is divided into two stages. In the first stage, visual processing quickly and spontaneously performs low-level feature extraction, including orientation, brightness, and color, from the visual input in various dimensions in a parallel manner. In the second stage, the visual processing will locate based on the features of the previous stage, generate a map of locations, and dynamically assemble the low-level features of each dimension of the activation area into high-level features. Generally speaking, essential areas attract the attention of the visual system more strongly. Wolfe believes that the attention mechanism uses not only the bottom-up information of the image but also the top-down information of the high-level visual organization structure [26], and the high-level information can effectively filter out a large amount of irrelevant information.

This work is inspired by the biological vision mechanism and proposes a novel attention learning method. We use traditional CNN to complete the extraction of local image features spontaneously. After that, we generate an attention map based on the low-level features of the previous stage to activate the spatial response of the feature, then calculate the attention vector based on the feature interdependence of the activation area to activate the channel response of the feature and finally complete the reorganization of the high-level features. The attention map and attention vector contain top-down information fed back to the current local features in the form of gating. We call this attention process a coarse-to-fine process, which is a hybrid domain attention mode that includes spatial-wise and channel-wise. Based on this method, we proposed a spatial and channel-wise coarse-to-fine attention network (SCOAT-Net) and used it to solve the segmentation task of COVID-19 lung opacification. As a summary, our contributions in this paper are threefold:

- We propose a novel coarse-to-fine attention network for segmentation of COVID-19 lung opacification from CT Images, which embedded spatial and channel-wise attention mechanism, and achieves the state-of-the-art performance (i.e., an average DCS of 0.8948).
- We use the self-attention method so that the neural network can generate attention maps without external ROI supervision. Furthermore, we use this method to understand the training process of the network by observing the areas that the network focuses on different stages and increasing the interpretability of the neural network.
- We verify the robustness and compatibility of the SCOAT-Net on different types of CT scans and confirm that it has specific data migration capability. Moreover, it can provide a quantitative assessment of pulmonary involvement difficult for radiologists and helps clinical follow-up of patient disease development and treatment response.

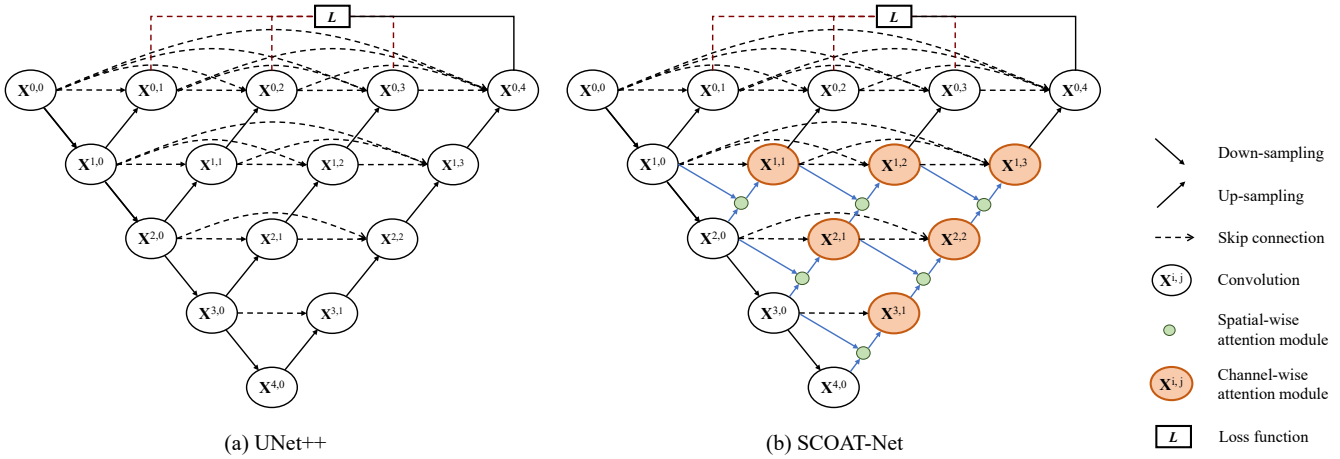
II. Related Works

A. Segmentation Networks

Nowadays, deep neural networks (DNN) have shown excellent performance for many automatic image segmentation tasks. Long et al. proposed the fully convolutional networks (FCN) [27], which uses the convolution neural network (CNN) in image semantic segmentation and achieves a breakthrough effect. Zhao et al. proposed the pyramid scene parsing network (PSPNet) [28], which introduces the global pyramid pooling into FCN to make the global and local information act on the prediction target together. DeeplabV3 [29] proposes the ASPP (atrous spatial pyramid pooling) module to make the segmentation model perform better on multi-scale objects. U-Net [13] was introduced by Ronneberger et al., based on the encoder-decoder structure that is widely used in medical image segmentation due to its excellent performance. It uses skip connections to connect the high-level low-resolution semantic feature map and the low-level high-resolution structural feature map of the encoder and decoder so that the network output has a better spatial resolution. Oktay et al. [21] proposed the attention gate model and applied it to the U-Net model, which improves the sensitivity and prediction accuracy of the model without increasing the calculation cost. UNet++ [30] uses a series of nested and dense skip paths to connect the sub-networks of encoder and decoder based on the U-NET framework, which further reduces the semantic relationship between encoder and decoder and achieves better performance in liver segmentation task.

B. Artificial Intelligence for COVID-19 based on CT

The segmentation of lung opacification based on CT images is an integral part of COVID-19 image processing, and there has been a series of related work. With using the lungs and pulmonary opacities manually segmented by experts as



(a) UNet++ (b) SCOAT-Net

Fig. 1. Comparison of UNet++ (a) and the proposed SCOAT-Net (b).

standards, Cao et al. [31] and Huang et al. [32] developed a CT image prediction model based on convolutional neural networks to monitor COVID-19 disease development, and quantification evaluation of lung involvement shows excellent potential. Some studies [33]–[36] trained segmentation models with CT and segmentation templates of abnormal lung cases, which can extract the areas related to lung diseases, making the learning process of pneumonia type classification easier in the next steps. The deep learning model relies on a large amount of data training, and it is impractical to collect a large amount of data with professional labels in a short time. [18]–[20] have maken attempt to solve this challenge from the perspectives of reducing manual delineation time, using noisy labels, and semi-supervised learning. VB-Net [18] has a perfect effect on the segmentation of COVID-19 infection regions. The mean percentage of infection (POI) estimation error for automatic segmentation and manual segmentation on the verification set is only 0.3%. In particular, it adopts a human-in-the-loop strategy to reduce the time of manual delineation significantly. Wang et al. [19] proposed noise-robust Dice loss and applied it in COPLE-Net, which surpasses other anti-noise training methods to learn COVID-19 pneumonia lesion segmentation in noisy labels. Inf-Net [20] uses a parallel partial decoder to aggregate high-level features and generate a global map to enhance the boundary area. They also uses a semi-supervised segmentation framework to achieve excellent performance in lung infection area segmentation.

C. Attention Mechanism

More and more attempts have been focused on the combination of deep learning and visual attention mechanism, which can be roughly divided into two categories: 1) external-attention and 2) self-attention. External-attention allows the network to learn to generate an attention map during the training process by giving a region of interest (ROI) supervision externally so that the region activated by the network can accurately diagnose the disease changes. [23], [37] had applied it to the diagnosis of COVID-19 and glaucoma, and the sensitivity was greatly improved. In contrast, self-attention does not rely on the supervision of external ROI, but exploits the intrinsic self-attention ability of CNN. The self-attention consists of two parts, among which spatial-wise attention [21], [38], [39] redistributes the network's attention at the pixel level of the feature map to achieve more precise location; channel-wise attention [40] redistributes the attention at channel level to instruct the network selecting practical features. [41] combines spatial and channel dimension attention with parallel mode to jointly guide network training, which captures rich contextual dependencies to address segmentation task. Chen et at. [42] proposed SCA-CNN in the task of image captioning, which incorporated spatial and channel-wise attentions. Zhang et al. [22] proposed an attention learning method with the higher layer feature as the attention mask of the lower layer feature, whice can achieve the best performance in skin lesion classification.

III. Method

UNet++ [30] is an excellent image segmentation network, which has achieved high-grade performance on medical image tasks. It contains dense connections, which make the contextual information of different scales closely related. However, although this complicated connection method improves the generalization ability of the model, it also causes information redundancy and weak convergence of the loss function on a small data set. Medical images have the characteristics of high complexity and noise, which cause the model overfitted when the training data is insufficient. The SCOAT-Net proposed in this work, redesigns the connection structure of UNet++ and introduces the attention learning mechanism. It extracts the spatial and channel features from coarse-to-fine with only a few added parameters and obtains a more accurate segmentation results.

A. Structure of the Lung Opacification Segmentation Network

Fig. 1 compares the basic structures of the UNet++ and the proposed SCOAT-Net. SCOAT-Net inherits the basic structure

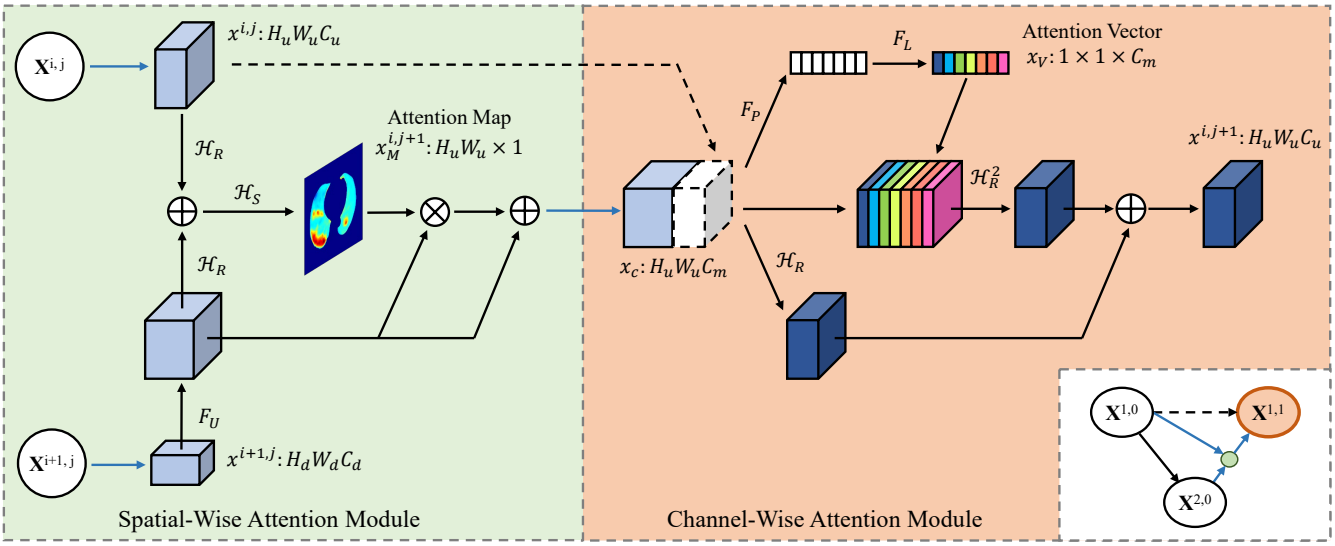


Fig. 2. Illustration of spatial-wise attention module and channel-wise attention module.

of UNet++, composing of encoder and decoder connected by skip connections. The encoder extracts the information of the semantic level of the image and gives relatively coarse location, which uses a max-pooling layer as a down-sampling module. The decoder reconstructs the segmentation template from semantic information. It uses U-shaped skip connections to receive the corresponding low-level features of the encoder and calculate the final segmentation result. The upsampling module of the decoder uses the bilinear interpolation layer instead of the deconvolution layer to improve the resolution of the feature map. This method dramatically reduces the number of parameters and saves the calculation cost, and has good performance on small-scale datasets.

We reconstruct the connection at the top of the network (except for the bottom layer $\mathbf{X}^{0,j}$) and introduce the attention module. This approach is to make the calculation of the attention mechanism act on the high-level semantic information and keep the bottom layer of the image detailed information as much as possible, makeing the segmentation result with more fined and higher resolution. The proposed attention module consists of two parts: the spatial-wise attention module and the channel-wise attention module.

We use the context feature maps with different resolutions as information of different dimensions to the spatial-wise attention module, as shown in the green circle of Fig. 1, which can combine all the multi-dimensional feature maps extracted by all the filters to calculate the attention map of the image and adjust the target area of the network adaptively. The output of the spatial-wise attention module is contacted with the feature maps of the same layer to enter the channel-wise attention module, as shown in the orange circle. The channel-wise attention module calculates the interdependence between the channels and adaptively recalibrates the information response of the channel. Besides, in each convolution module, we use the residual block the training of our network.

B. Spatial-Wise Attention

The proposed spatial-wise attention module emphasizes attention at the pixel level, making the network pay attention to the key formation and ignore irrelevant information. Normally, in a convolutional neural network, the features extracted by the network change from simple low-level features to complex high-level features with the deepening of the convolutional layers. We except that when calculating the attention map, we can not only use the information of single-layer features but also combine the upper and lower features of different resolutions. The final output of this module is expressed as $x_s \in \mathbb{R}^{H_u \times W_u \times C_u}$, which is given by (1) and (2):

$$x_M^{i,j+1} = \mathcal{H}_S\left(\mathcal{H}_R\left(x^{i,j}\right) + \mathcal{H}_R\left(F_U(x^{i+1,j})\right)\right) \quad (1)$$

$$x_s = (1 + x_M^{i,j+1}) \cdot F_U(x^{i+1,j}) \quad (2)$$

where function $\mathcal{H}_R(\cdot)$ stands for the convolution of size 1×1 followed by a batch normalization and a ReLU, used for feature integration. $\mathcal{H}_S(\cdot)$ denotes the convolution of size 1×1 followed by a batch normalization and a sigmoid activation function, used for feature integration and generation of the attention maps. $F_U(\cdot)$ is the up-sampling operation, with bilinear interpolation function. The input of this module is composed of the upper layer feature $x^{i,j} \in \mathbb{R}^{H_u \times W_u \times C_u}$ and the lower layer feature $x^{i+1,j} \in \mathbb{R}^{H_d \times W_d \times C_d}$, where $x^{i,j}$ represents the output of each convolution module $\mathbf{X}^{i,j}$. $x_M \in \mathbb{R}^{H_u \times W_u \times 1}$ is the attention map generated by this module, which uses the saliency information in the spatial position to weigh the input features to complete the redistribution of the feature attention at the pixel level. The attention map generated by the sigmoid function is normalized between 0 and 1, and the output response will be weakened after point multiplication with the current feature map. Nested structure uses of this method will lead to over-fitting or the degradation of model performance caused by the gradient's disappearance. To improve this phenomenon, inspired by the ResNet, we add the original features $x^{i+1,j}$ after weighting them by the $x_M^{i,j+1}$,

as shown in (2). The final output x_s is sent to the next channel-wise attention module.

C. Channel-Wise Attention

The input $x_c \in \mathbb{R}^{H_u \times W_u \times C_m}$ of the proposed channel-wise attention module is obtained by concatenating the spatial-wise attention module's output x_s with the feature map of the same layer, as in (3):

$$x_c = \left[\left[x^{i,k} \right]_{k=0}^{j-1}, x_s \right] \tag{3}$$

where $[\cdot]$ means the concatenation. $x_g \in \mathbb{R}^{1\times1\times C_m}$ is the channel-wise statistical information calculated by x_c through a global average pooling layer, as in (4), which can reflect the response degree on each feature map.

$$x_g = F_P(x_c) = \frac{1}{H_u \times W_u} \sum_{i=1}^{H_u} \sum_{j=1}^{W_u} x_c(i,j) \tag{4}$$

On one hand, we want the module to adaptively learn the feature channels that require more attention. On the other hand, we want it to learn the interdependence between channels. Inspired by the SENet [40], we pass x_g through two fully connected (FC) layers with parameters ω_1 and ω_2 to obtain the attention vector $x_V \in \mathbb{R}^{1\times1\times C_m}$ of the channel, as in (5):

$$x_V = F_L(x_g) = \sigma\left(\omega_2 \rho\left(\omega_1 x_g\right)\right) \tag{5}$$

where $\rho(\cdot)$ refers to the ReLU activation function, and $\sigma(\cdot)$ refers to the sigmoid activation function. A structure containing two fully connected layers is adopted here, which reduces the complexity and improves the generalization ability of the model. The fully connected layer of parameter $\omega_1 \in \mathbb{R}^{\frac{C_m}{r}\times C_m}$ reduces the feature channels' dimension with reduction ratio r ($r = 16$ in this experiment). In contrast, the fully connected layer of parameter $\omega_2 \in \mathbb{R}^{C_m \times \frac{C_m}{r}}$ recombines the feature channels to incease its dimension to the C_m. The attention vector x_V finally weights the input feature map x_c, and after the convolution operation completes the feature extraction, it is added to itself to obtain the final output $x^{i,j+1} \in \mathbb{R}^{H_u \times W_u \times C_u}$, as in (6):

$$x^{i,j+1} = \mathcal{H}_R^2(x_V \cdot x_c) + \mathcal{H}_R(x_c) \tag{6}$$

where $\mathcal{H}_R^2(\cdot)$ represents the two-layer convolution for feature extraction.

D. Loss Function

SCOAT-Net has a deep supervision strategy, which can use any one of the segmentation branch outputs ($x^{0,j}, j \in 1,2,3,4$) to calculate the loss or use the output of all branches to calculate the average of the loss. The choice depends on the tasks and data. By combining binary cross-entropy (BCE) loss and dice coefficient loss [43], we use a hybrid loss function for segmentation as follows:

$$\begin{aligned}\mathcal{L}_{seg} &= \mathcal{L}_{bce} + \alpha \times \mathcal{L}_{dice} \\ &= -\frac{1}{N}\sum_{b=1}^{N}\left(Y_b \cdot \log(\sigma(\hat{Y_b})) + (1-Y_b)\cdot log(\sigma(1-\hat{Y_b}))\right) \\ &\quad - \frac{2\alpha \times \boldsymbol{Y}\cdot\hat{\boldsymbol{Y}}}{\boldsymbol{Y}^2 + \hat{\boldsymbol{Y}}^2}\end{aligned} \tag{7}$$

where $\boldsymbol{Y} = \{Y_1, Y_2, \cdots, Y_b\}$ denotes the ground truths, $\hat{\boldsymbol{Y}}$ denotes the predicted probabilities, N indicates the batch size, and $\sigma(\cdot)$ corresponds to the sigmoid activation function. This hybrid loss includes pixel-level and batch-level information, which helps the network parameters to be better optimized.

E. Evaluation Metrics

To evaluate the performance of lung opacification segmentation, we measure the Dice similarity coefficient (DSC), sensitivity (SEN), positive predicted value (PPV), volume accuracy (VA), regional level precision (RLP), and regional level recall (RLR) betweent the segmentation results and the ground truth, which are defined as follows.

$$\begin{aligned} DSC &= \frac{2|V_a \cap V_b|}{|V_a| + |V_b|},\ SEN = \frac{|V_a \cap V_b|}{|V_b|}, \\ PPV &= \frac{|V_a \cap V_b|}{|V_a|},\ VA = 1 - \frac{2abs(|V_a| - |V_b|)}{|V_a| + |V_b|} \end{aligned} \tag{8}$$

where V_a and V_b refer to the segmented volumes by the model and the ground truth, respectively. In addition to the above voxel-level evaluation indicators, we also design the regional-level evaluation indicators RLP and RLR, as in (9):

$$RLP = \frac{N_p}{N_a},\ RLR = \frac{N_t}{N_b} \tag{9}$$

N_a denotes the total number of connected regions of the model prediction result, N_p denotes the number of real opacity regions predicted by the model, N_b denotes the total number of real opacitiy regions, and N_t denotes the number of real opacity regions predicted by the model. If the center of the connected area predicted by the model is in the real opacity regions, then we accept that the predicted connected area is correct. We calculate the center of the connected area as:

$$u = \arg\min_i \max_j \|u_i - v_j\|,\ (u_i \in U, v_j \in V) \tag{10}$$

where U represents the point set of a single connected area of the prediction result, and V represents the point set of its edge.

IV. Experiment and Results

A. Data and Implementation

This study and procedures was appreoved by the local ethics committees. All methods were performed in accordance with the relevant guidelines and regulations. Written informed consent from the study patients was not required. The data contains 19 lung CT scans of COVID-19 patients scanned using SOMATOM Definition AS, and the 1117 lung opacification segmentation delineated by radiologists on the single-slice CT. Besides, we prepared a total of 8 lung CT scans of 2 patients scanned at different times, using SOMATOM go.Top, which is used to test the compatibility of our model in different device types. We performed 5-fold cross-validation to test the results. The input images are single-layer CT images, which are in the size of 512×512 pixels to ensure the high resolution of the result, normalized before sent to the network. The sketch templates of the radiologists serve as the ground truth, which

used to calculate the loss function with the final output of the network. We use the gradient descent algorithm with Adam [44] to optimize the loss function that updates the network parameters. The learning rate is set to 0.01, which is multiplied by 0.1 after every ten epoch decay. When the iterative result converges, we adjust the learning rate to 0.001 for training again. The learning rate decay strategy remains unchanged, and the iteration is set to 50 times. The final results of training in this warm-up [45] method will be slightly improved. All experiments were conducted on a NVIDIA RTX GPU, and the proposed SCOAT-Net were implemented based on Pytorch framework.

TABLE I
QUANTITATIVE EVALUATION OF SCOAT-NET WITH DEFFERENT LOSS FUNCTIONS FOR LUNG OPACIFICATION SEGMENTATION.

Loss functions	Results (%)					
	DSC	SEN	PPV	VA	RLP	RLR
MAE [46]	85.74	84.97	83.52	91.78	82.49	84.77
IOU [47]	88.10	86.36	90.47	94.69	92.28	88.17
BCE	88.53	86.85	91.43	94.70	92.47	89.93
Dice [43]	86.79	89.24	85.38	90.78	86.77	92.30
Focal [48]	87.73	87.28	87.93	93.86	88.18	88.96
BCE-Dice ($\alpha = 0.5$)	89.48	88.74	90.64	90.64	93.04	91.18

B. Results on Lung Opacification Segmentation

This experiment aims to evaluate the performance of our proposed SCOAT-Net with different loss functions for lung opacification segmentation. We tried six loss functions: MAE [46], IOU [47], BCE, Dice [43], Focal [48], and BCE-Dice to train the proposed network, with the same strategy and hyper-parameters, and the quantitative comparison is listed in Table I. It is evident that IOU, BCE, and Focal have excellent segmentation performance, and their DSCs are at the higher levels. Among them, Focal is higher than IOU and BCE on SEN, and slightly inferior on PPV and RLP. It is worth noting that Dice has a more significant performance on SEN and RLR. The results can predict the entire opacity area better, but it also causes PPV and RLP performance declined because of predicting more false-positive areas. The hybrid loss function combined BCE and Dice with parameter α (We set $\alpha = 0.5$ in the experiments.) produces the best results. Except for SEN and RLR, which are slightly lower than Dice, the other indicators have achieved first place. The box plot shown in Fig. 3 demonstrates the performance of our proposed network with the BCE-Dice loss function. It is observed that in 19 cases, the model we proposed has a very high performance. The medians of DSC, SEN, and PPV are all higher than 0.9, and the medians of RLP and PLR are higher than 0.95, even one or two cases did not achieve excellent results.

C. Comparison of Different Networks

We compared our proposed SCOAT-Net with other popular segmentation algorithms for lung opacification segmentation. The BCE-Dice loss function was used to train these networks. The quantitative evaluation of these networks was calculated

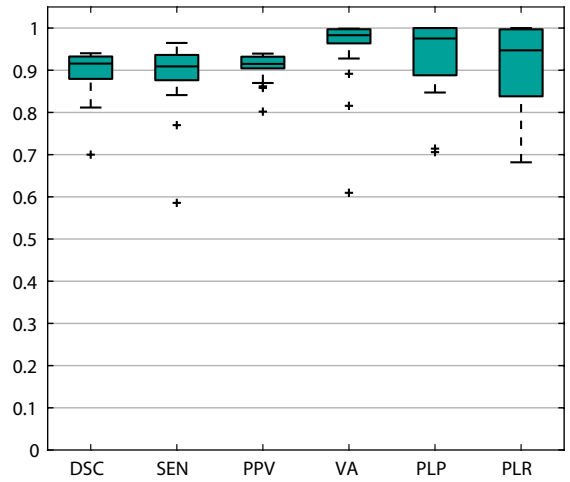


Fig. 3. The segmentation performances of SCOAT-Net with BCE-Dice loss function.

TABLE II
QUANTITATIVE EVALUATION OF DEFFERENT NETWORKS FOR LUNG OPACIFICATION SEGMENTATION. THE BCE-DICE LOSS WAS USED FOR TRAINING.

Methods	Results (%)					
	DSC	SEN	PPV	VA	RLP	RLR
FCN [27]	82.21	84.07	80.78	93.16	81.94	82.18
DeepLabV3 [29]	84.81	86.32	83.62	93.35	87.80	87.95
PSPNet [28]	85.08	84.95	85.84	92.40	90.81	86.07
Attention UNet [21]	85.64	84.87	86.69	86.69	91.70	86.49
U-Net [13]	84.72	82.23	87.90	90.46	91.48	87.03
UNet++ [30]	84.82	83.55	86.35	94.09	89.05	84.75
SCOAT-Net	89.48	88.74	90.64	94.97	93.04	91.18

by cross-validation, as shown in Table II. The performance of FCN segmentation results on various indicators is not very good, which demonstrates it is challenging to obtain excellent segmentation results using the neural networks that only contain convolutional structures. U-Net, which has excellent performance on many medical image segmentation tasks, has the excellent RLP but the lowest SEN. Although most of the predicted regions are correct, the voxel prediction of opacity regions cannot capture completely. Compared with U-Net, which has a more complex structure and more connections, UNet++ has slightly improved performance on DSC, SEN, and VA, but it has a significant drop on RLP and RLR, which shows that its dense connection improves the model's generality. However, it is difficult to achieve excellent results on the reletively small dataset used in this work.

Our proposed SCOAT-Net achieves the best performance among the compared networks. It can more effectively identify and segment the pulmonary opacities by using spatial and channel-wise attention modules. Fig. 4 shows a visual comparison of the results of each network. On four cases, the results show that SCOAT-Net has the best segmentation performance, which can not only effectively hit the target opacity region but also produce the least difference between the segmentation area and the ground truth. However, there are also unsatisfactory segmentation results, as shown in the

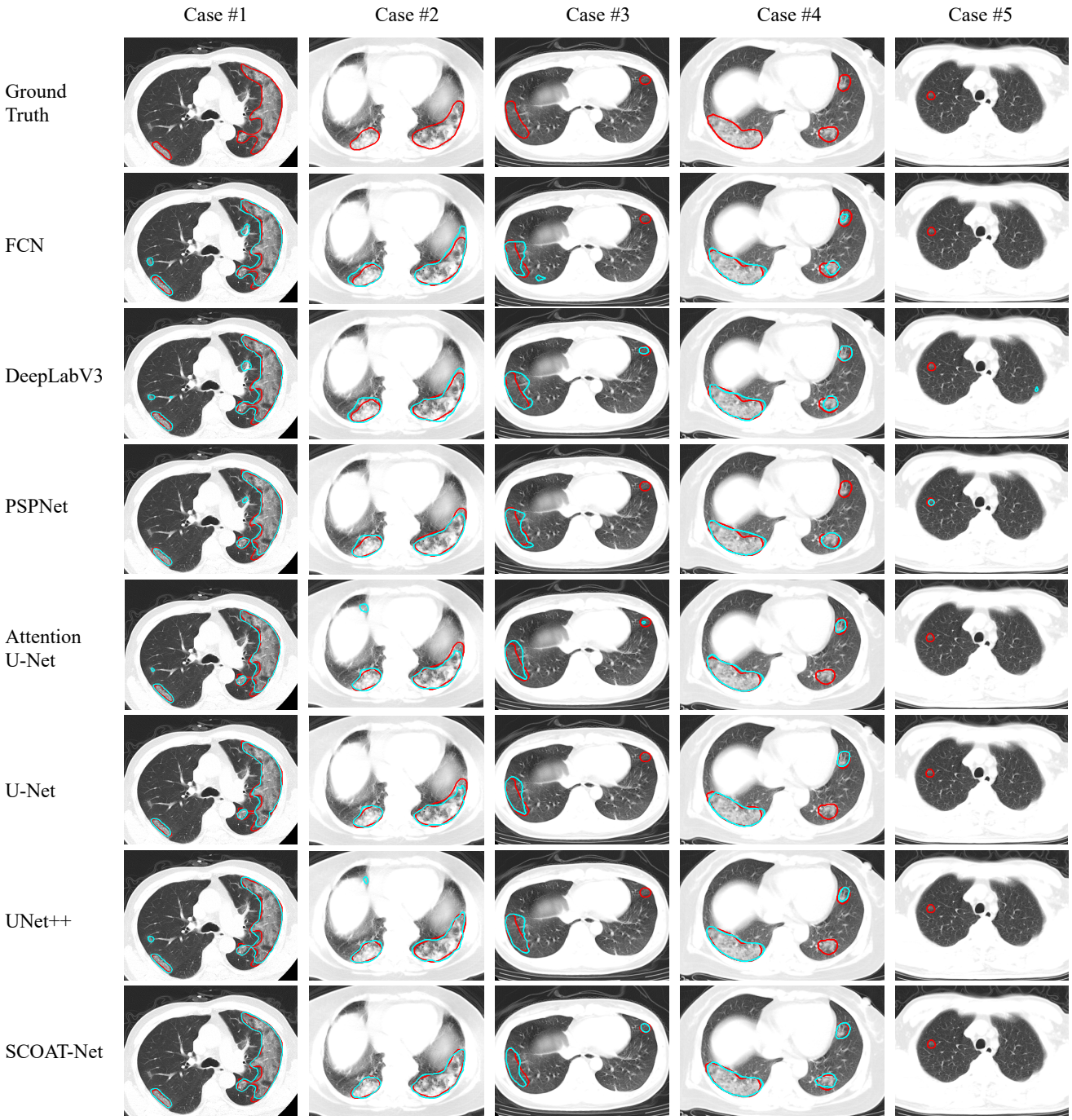


Fig. 4. Visual comparison of segmentation performance of different model trained with BCE-Dice loss function. The red curves represent the ground truth, and the cyan curves represent the results of the model.

right-most column of Fig .4. Most of the models, including our model, failed to predict this tiny opacity region, but PSPNet made valid prediction. This indicates that PSPNet, which adopts multi-scale feature fusion, performs excellently in segmenting regions of different scales.

D. Effectiveness of the Attention Module

In this experiment, we verified the performance of the attention module on the lung opacification segmentation task. Our SCOAT-Net uses a total of 6 spatial-wise attention modules, as shown in the green circle in Fig .1. These modules can adaptively generate attention maps with the focused area information of the network. The early stage of our network is defined as the position that closes to the input and passes fewer convolution layers. The later stage is defined as the position that closes to the output and passes more convolution layers. We selected three different stages of attention maps for display, and the order from the early stage to the late stage is $x_M^{1,1}$, $x_M^{2,2}$, and $x_M^{1,3}$, as shown in Fig .5. For better display, we only

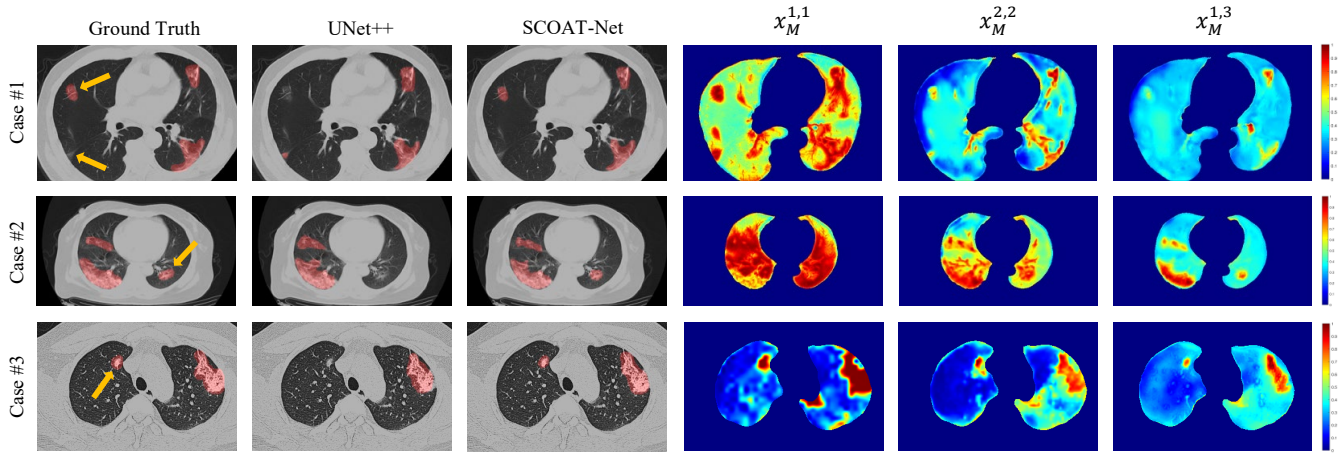


Fig. 5. Visualization of the segmention results and attention maps of our methods on three COVID-19 cases. The red area is the lung opacification segmentation of the ground truth and the models of UNet++ and our SCOAT-Net, and yellow arrows highlight the local difference of the segmentation results.

show the lung area. We can see that our SCOAT-Net has an excellent performance on lung opacification recognition than UNet++. For example, in the first case, UNet++ identified the interlobular fissure (the yellow arrow area in the lower-left corner) with a specific shape and structure as an opacity region, but our SCOAT-Net did not misidentify it. From the attention map of this case, we can see that $x_M^{1,1}$ focuses on all the salient areas of the lungs, basically covering all the structures of the lung. Furthermore, $x_M^{2,2}$ reduces largely the significant areas, and the attention of the network is more concentrated on restricted regions. By $x_M^{1,3}$, the interlobular fissure area misidentified in the early stage has no longer received the core attention. Besides, for the opacity region that UNet++ did not recognize (the region indicated by the yellow arrow), SCOAT-Net has adequately identified the target area, and on all the attention maps, much attention focus on the target area.As the training phase progresses, the attented regions of the SCOAT-Net gradually become less. The attention module we designed not only effectively weights the feature map, but also further helps us understand the training process of the neural network, which improves its interpretability.

Furthermore, we also introduced the attention module from other studies into UNet++ and compared the results with that of our SCOAT-Net, as shown in Table III. AttentionV1 uses the attention module of residual attention network [38], AttentionV2 imitates the connection structure of Attention UNet [21], and AttentionV3 uses the pyramid attention module of Wang et al. [39]. Compared with the baseline UNet++, all the networks have obtained the significantly improved DSC and RLR. SCOAT-Net and AttentionV2 have outstanding performance in SEN, and SCOAT-Net, AttentionV2, and AttentionV3 have also significantly improved in the RLP. The results show that the attention module can improve the segmentation performance while only increasing a few parameters of the network, especially for the recognition of the target area.

TABLE III
QUANTITATIVE EVALUATION OF DEFFERENT ATTENTION MODULE FOR SEGMENTATION. THE BASELINE NETWORK IS UNET++.

Methods	Params	Results (%)					
		DSC	SEN	PPV	VE	RLP	RLR
UNet++	35.02M	84.82	83.55	86.35	94.09	89.05	84.75
AttentionV1	53.43M	86.22	85.56	87.44	92.96	90.53	87.57
AttentionV2	37.63M	88.01	87.14	89.28	94.70	93.28	90.57
AttentionV3	35.05M	86.91	84.76	89.55	92.38	93.22	89.55
SCOAT-Net	39.08M	89.48	88.74	90.64	94.97	93.04	91.18

E. Validation on Other Data

We prepared a total of 8 lung CT scans of 2 patients scanned at different time, of which the type of CT device is SOMATOM go.Top different from in training, which is used to test the robustness and compatibility of the proposed SCOAT-Net. We show the lung CT of two cases under treatment in Fig .6. COVID-19 is clinically divided into four stages [49]: early stage, progressive stage, peak stage, and absorption stage. The clinical report of the first case shows that it was in the absorption stage at all four time points. From the result of our model, we can see that on both of the axial unenhanced or coronal reconstructions CT images, the opacity regions are significantly reduced, which are further verified by the Lung opacification volume (LOV) displayed on the lower right corners of the coronal images. The clinical report of the second case shows that the first time point is the early stage, the second time point was the progressive stage, and the third and fourth time points are the absorption stage. Our calculated LOV is highest at the second time point, and there is a significant decrease in the third time point, which also matches the diagnosis report of the patient. In summary, our proposed SCOAT-Net on cross-modal CT scans is verified, proving that it has better robustness and compatibility. It can provide an objective assessment of pulmonary involvement and therapy response in COVID-19.

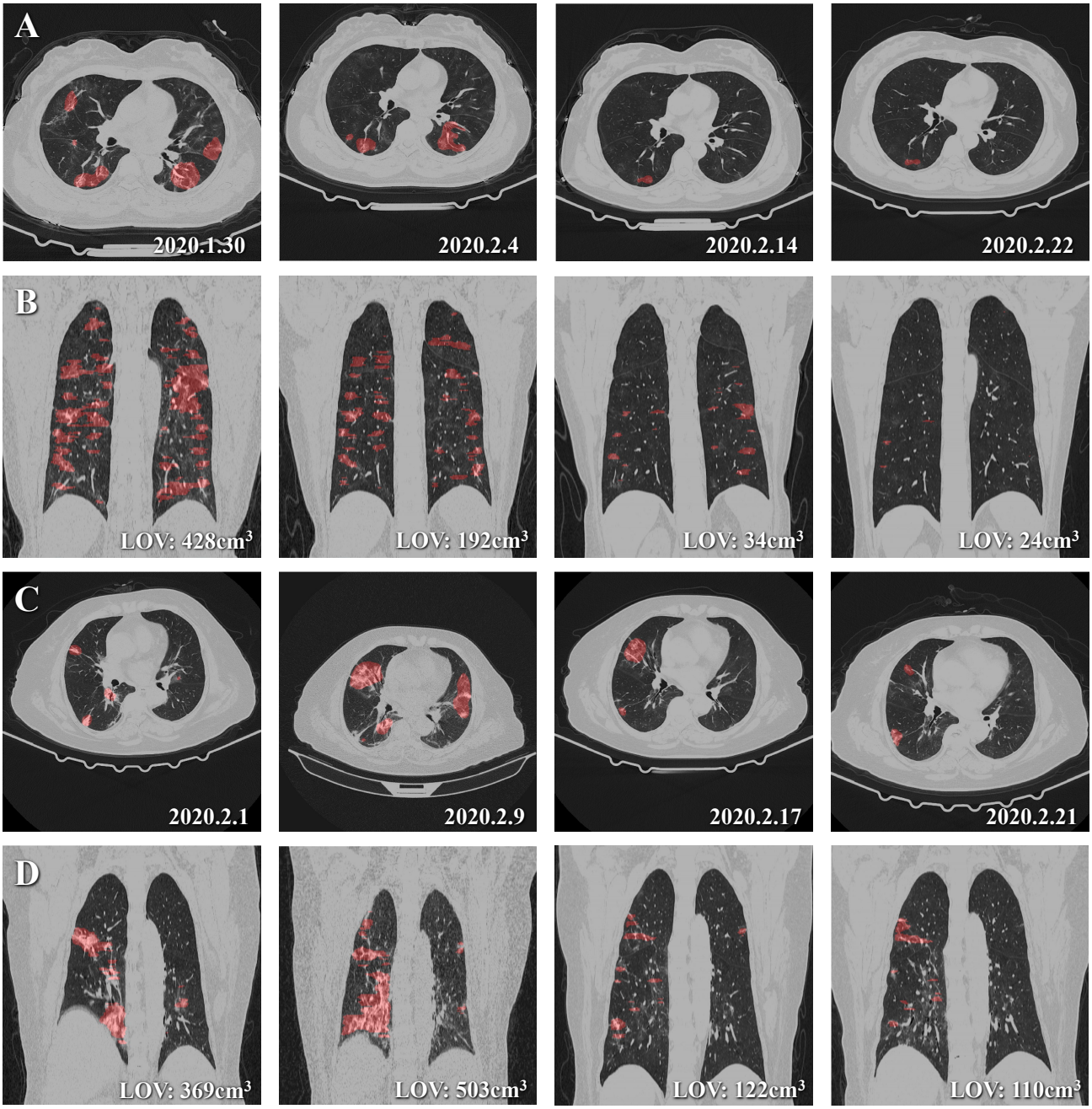


Fig. 6. Qualitative evaluation of the results of SCOAT-Net on the data from other type of CT scan. A and B are the evolution of one COVID-19 case during the 24-days treatment period, C and D is the evolution of another case during the 21-days treatment period. A and C are axial unenhanced chest CT images at four time points (dates are annotated at the lower-right corner of each panel), B and D are the coronal reconstructions at the same time points. The segmentation of pulmonary opacities derived from SCOAT-Net displayed in red, and the volumetric assessment of our results (i.e., Lung opacification volume (LOV)) is annotated at the lower-right corners of the images of B and C.

V. Discussion and Conclusion

CNN has been widely used in various medical image segmentation tasks due to its excellent performance [13], [21], [30], [47]. Some networks have been improved from the perspective of connection structure (e.g., U-Net), and others have been improved from the perspective of combining multi-scale features (e.g., PSPNet). These improvements have enhanced the expression ability of the models to a certain extent. However, due to the particularity of medical image related tasks, only a small amount of data can be obtained, making it impossible to converge when training conventional deep neural networks effectively, which is a common problem. In addition to augmenting the data [50], some works show that attentional mechanisms can be more effective in enhancing the generalization capacity of models.

We imitated the biological vision mechanism, tightly integrated the attention mechanism into the CNN training process, and proposed a spatial and channel-wise coarse-to-fine

attention network. We applied it to the segmentation task of lung opacification segmentation in COVID, and achieved better performance than state-of-the-art CNNs of the image segmentation, as shown in Table 2. Furthermore, we compared the influence of four different attention modules in other models and ours on this task. The network incorporating the attention module has improved performance to varying degrees compared to the baseline network. It is worth mentioning that the attention module we proposed generates a series of attention maps. We can observe the changes of the focused regions at different stages, which contributes to the interpretability of the neural network.

Also, we verified the robustness and compatibility of our model on different types of CT equipments and confirmed that it has excellent data migration capability. Our model can accuracy segment the lung opacity regions on CT images at different time-points during the treatment. It provides a quantitative assessment of pulmonary involvement that is difficult for radiologists, which helps clinical follow-up of patient disease development and treatment response.

However, our model still has shortcomings. In the failure case shown in the rightmost column of Figure 4, only PSP-Net, which incorporates multi-scale features, can identify tiny opacity regions. This suggests that we can continue to enhance the model's recognition of targets of different scales, by using multi-scale feature fusion or cascading convolution in different receptive field sizes.

REFERENCES

[1] J. T. Wu, K. Leung, and G. M. Leung, "Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study," *The Lancet*, vol. 395, no. 10225, pp. 689–697, 2020.

[2] Z. Wu and J. M. McGoogan, "Characteristics of and important lessons from the coronavirus disease 2019 (covid-19) outbreak in china: Summary of a report of 72 314 cases from the chinese center for disease control and prevention," *JAMA*, vol. 323, no. 13, pp. 1239–1242, 04 2020.

[3] H. Shi, X. Han, N. Jiang, Y. Cao, O. Alwalid, J. Gu, Y. Fan, and C. Zheng, "Radiological findings from 81 patients with covid-19 pneumonia in wuhan, china: a descriptive study," *Lancet Infectious Diseases*, 2020.

[4] Z. Xu, L. Shi, Y. Wang, J. Zhang, L. Huang, C. Zhang, S. Liu, P. Zhao, H. Liu, L. Zhu *et al.*, "Pathological findings of covid-19 associated with acute respiratory distress syndrome," *The Lancet Respiratory Medicine*, 2020.

[5] "Weekly operational update coronavirus disease 2019 (covid-19)," [EB/OL], 2020, https://www.who.int/docs/default-source/coronaviruse/weekly-updates/wou-9-september-2020-cleared-14092020.pdf?sfvrsn=68120013_2.

[6] Z. Y. Zu, Jiang, P. P. Xu, W. Chen, Q. Q. Ni, G. Lu, and L. J. Zhang, "Coronavirus disease 2019 (covid-19): A perspective from china," *Radiology*, pp. 200 490–200 490, 2020.

[7] Y. Fang, H. Zhang, J. Xie, M. Lin, L. Ying, P. Pang, and W. Ji, "Sensitivity of chest ct for covid-19: Comparison to rt-pcr," *Radiology*, pp. 200 432–200 432, 2020.

[8] J. F. W. Chan, S. Yuan, K. Kok, K. K. W. To, H. Chu, J. Yang, F. Xing, J. Liu, C. C. Yip, R. W. S. Poon *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.

[9] T. Ai, Z. Yang, H. Hou, C. Zhan, C. Chen, W. Lv, Q. Tao, Z. Sun, and L. Xia, "Correlation of chest ct and rt-pcr testing in coronavirus disease 2019 (covid-19) in china: A report of 1014 cases," *Radiology*, pp. 200 642–200 642, 2020.

[10] M. Chung, A. Bernheim, X. Mei, N. Zhang, M. Huang, X. Zeng, J. Cui, W. Xu, Y. Yang, Z. A. Fayad *et al.*, "Ct imaging features of 2019 novel coronavirus (2019-ncov)," *Radiology*, vol. 295, no. 1, pp. 202–207, 2020.

[11] Z. Ye, Y. Zhang, Y. Wang, Z. Huang, and B. Song, "Chest ct manifestations of new coronavirus disease 2019 (covid-19): a pictorial review," *European Radiology*, pp. 1–9, 2020.

[12] A. Esteva, B. Kuprel, R. A. Novoa, J. M. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[14] P. Kickingereder, F. Isensee, I. Tursunova, J. Petersen, U. Neuberger, D. Bonekamp, G. Brugnara, M. Schell, T. Kessler, M. Foltyn *et al.*, "Automated quantitative tumour response assessment of mri in neuro-oncology with artificial neural networks: a multicentre, retrospective study," *Lancet Oncology*, vol. 20, no. 5, pp. 728–740, 2019.

[15] Y. Lecun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[16] D. S. Kermany, M. H. Goldbaum, W. Cai, C. C. S. Valentim, H. Liang, S. L. Baxter, A. Mckeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[17] Y. Xie, Y. Xia, J. Zhang, Y. Song, D. Feng, M. J. Fulham, and W. Cai, "Knowledge-based collaborative deep learning for benign-malignant lung nodule classification on chest ct," *IEEE Transactions on Medical Imaging*, vol. 38, no. 4, pp. 991–1004, 2019.

[18] F. Shan, Y. Gao, J. Wang, W. Shi, N. Shi, M. Han, Z. Xue, and Y. Shi, "Lung infection quantification of covid-19 in ct images with deep learning." *arXiv: Computer Vision and Pattern Recognition*, 2020.

[19] G. Wang, X. Liu, C. Li, Z. Xu, J. Ruan, H. Zhu, T. Meng, K. Li, N. Huang, and S. Zhang, "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.

[20] D. Fan, T. Zhou, G. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, and L. 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.

[21] O. Oktay, J. Schlemper, L. L. Folgoc, M. C. H. Lee, M. P. Heinrich, K. Misawa, K. Mori, S. Mcdonagh, N. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv: Computer Vision and Pattern Recognition*, 2018.

[22] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention residual learning for skin lesion classification," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2092–2103, 2019.

[23] X. Ouyang, J. Huo, L. Xia, F. Shan, J. Liu, Z. Mo, F. Yan, Z. Ding, Q. Yang, B. Song, F. Shi, H. Yuan, Y. Wei, X. Cao, Y. Gao, D. Wu, Q. Wang, and D. Shen, "Dual-sampling attention network for diagnosis of covid-19 from community acquired pneumonia," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2595–2605, 2020.

[24] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention." *Vision Research*, vol. 40, no. 10, pp. 1489–1506, 2000.

[25] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[26] J. M. Wolfe, M. L.-H. Võ, K. K. Evans, and M. R. Greene, "Visual search in scenes involves selective and nonselective pathways," *Trends in cognitive sciences*, vol. 15, no. 2, pp. 77–84, 2011.

[27] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *arXiv: Computer Vision and Pattern Recognition*, 2014.

[28] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," *arXiv: Computer Vision and Pattern Recognition*, 2016.

[29] L. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv: Computer Vision and Pattern Recognition*, 2017.

[30] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham: Springer International Publishing, 2018, pp. 3–11.

[31] Y. Cao, Z. Xu, J. Feng, C. Jin, X. Han, H. Wu, and H. Shi, "Longitudinal assessment of covid-19 using a deep learning–based quantitative ct pipeline: Illustration of two cases," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 2, p. e200082, 2020.

[32] L. Huang, R. Han, T. Ai, P. Yu, H. Kang, Q. Tao, and L. Xia, "Serial quantitative chest ct assessment of covid-19: Deep-learning approach," *Radiology: Cardiothoracic Imaging*, vol. 2, no. 2, p. e200075, 2020.

[33] O. Gozes, M. Frid-Adar, H. Greenspan, P. D. Browning, H. Zhang, W. Ji, A. Bernheim, and E. Siegel, "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.

[34] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song *et al.*, "Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct," *Radiology*, 2020.

[35] J. Chen, L. Wu, J. Zhang, L. Zhang, D. Gong, Y. Zhao, S. Hu, Y. Wang, X. Hu, B. Zheng *et al.*, "Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study," *MedRxiv*, 2020.

[36] S. Jin, B. Wang, H. Xu, C. Luo, L. Wei, W. Zhao, X. Hou, W. Ma, Z. Xu, Z. Zheng *et al.*, "Ai-assisted ct imaging analysis for covid-19 screening: Building and deploying a medical ai system in four weeks," *medRxiv*, 2020.

[37] L. Li, M. Xu, X. Wang, L. Jiang, and H. Liu, "Attention based glaucoma detection: A large-scale database and cnn model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 571–10 580.

[38] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.

[39] W. Wang, S. Zhao, J. Shen, S. C. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1448–1457.

[40] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[41] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.

[42] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," pp. 6298–6306, 2017.

[43] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv: Learning*, 2014.

[45] A. Gotmare, N. S. Keskar, C. Xiong, and R. Socher, "A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation," *arXiv: Learning*, 2018.

[46] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance," *Climate research*, vol. 30, no. 1, pp. 79–82, 2005.

[47] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "Unet 3+: A full-scale connected unet for medical image segmentation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1055–1059.

[48] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[49] H. Li, S. Liu, H. Xu, and J. Cheng, "Guideline for medical imaging in auxiliary diagnosis of coronavirus disease 2019," *Chin J Med Imaging Technol*, vol. 36, no. 3, pp. 321–331, 2020.

[50] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca, "Data augmentation using learned transformations for one-shot medical image segmentation," *arXiv: Computer Vision and Pattern Recognition*, 2019.