FSS-2019-nCov: A Deep Learning Architecture for Semi-supervised Few-Shot Segmentation of COVID-19 Infection

Mohamed Abdel-Basset[1], Victor Chang[2], Hossam Hawash[1], Ripon K. Chakrabortty[3] and Michael Ryan[3]

[1]Faculty of Computers and Informatics, Zagazig University, Zagazig, Sharqiyah, 44519, Egypt.
Emails: *mohamedbasset@zu.edu.eg; hossammoh@zu.edu.eg*

[2]School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, UK
Email: victorchang.research@gmail.com/V.Chang@tees.ac.uk

[3]Capability Systems Centre, School of Engineering and IT, UNSW Canberra; Australia
Emails: *r.chakrabortty@adfa.edu.au; m.ryan@adfa.edu.au*

Abstract

The newly discovered coronavirus (COVID-19) pneumonia is providing major challenges to research in terms of diagnosis and disease quantification. Deep-learning (DL) techniques allow extremely precise image segmentation; yet, they necessitate huge volumes of manually labeled data to be trained in a supervised manner. Few-Shot Learning (FSL) paradigms tackle this issue by learning a novel category from a small number of annotated instances. We present an innovative semi-supervised few-shot segmentation (FSS) approach for efficient segmentation of 2019-nCov infection (FSS-2019-nCov) from only a few amounts of annotated lung CT scans. The key challenge of this study is to provide accurate segmentation of COVID-19 infection from a limited number of annotated instances. For that purpose, we propose a novel dual-path deep-learning architecture for FSS. Every path contains encoder-decoder (E-D) architecture to extract high-level information while maintaining the channel information of COVID-19 CT slices. The E-D architecture primarily consists of three main modules: a feature encoder module, a context enrichment (CE) module, and a feature decoder module. We utilize the pre-trained ResNet34 as an encoder backbone for feature extraction. The CE module is designated by a newly introduced proposed Smoothed Atrous Convolution (SAC) block and Multi-scale Pyramid Pooling (MPP) block. The conditioner path takes the pairs of CT images and their labels as input and produces a relevant knowledge representation that is transferred to the segmentation path to be used to segment the new images. To enable effective collaboration between both paths, we propose an adaptive recombination and recalibration (RR) module that permits intensive knowledge exchange between paths with a trivial increase in computational complexity. The model is extended to multi-class labeling for various types of lung infections. This contribution overcomes the limitation of the lack of large numbers of COVID-19 CT scans. It also provides a general framework for lung disease diagnosis in limited data situations.

Keywords: Deep Learning; Few-Shot Segmentation; COVID-19; Context Fusion; CT images

I. INTRODUCTION

In December 2019, a global health crisis began with the spread of the novel Coronaviridae species called severe acute respiratory syndrome coronavirus 2 (SARS-COV-2)--specifically, the novel Coronavirus Disease (COVID-19) [1]. Over the last few months, the CSSE at Johns Hopkins University has reported 4,733,349 infections and 313,384 deaths in 180 countries around the world[1] (online access: 17 May). The reverse-transcription polymerase chain reaction (RT-PCR) is regarded as the major means for inspecting COVID-19 infection. However, the lack of equipment and the restrictions of appropriate testing settings limit rapid and precise screening. Additionally, the RT-PCR test has also been shown to have high false-negative rates [2]. Radiological imaging methods (such as X-ray and computed tomography (CT)) provide a significant supplement to RT-PCR tests and have shown their efficiency in lung disease diagnosis and quantification [3]. Moreover, several studies show that chest CT analysis results in higher performance (greater sensitivity) in COVID-19 detection compared to RT-PCR [4]. In comparison to X-rays, CT screening has the advantages of a three-dimensional representation of the patient's lung. Recent studies [5] indicate that the distinctive infection indication of ground-glass opacity (GGO) and consolidation could be detected from CT scans. The GGO was defined as hazy growing lung attenuation with the conservation of bronchial and vascular margins. In contrast, the consolidation was identified as opacification with obscuration boundaries of bowls and airway walls [6]. Therefore, the qualitative assessment of contagion and longitudinal variations in CT images could provide beneficial and substantial information about COVID-19. However, the manual projection of lung infections is laborious and time-consuming and the accuracy of infection annotation depends heavily on the knowledge and experience of the radiologist. There is, therefore, a need for automatic and accurate segmentation techniques that enable rapid screening of COVID-19.

Recently, a wide variety of deep-learning approaches has been used for semantic image segmentation. Among them, fully convolutional neural networks (F-CNNs) have shown superior performance on both traditional and medical images [7-13]. Notwithstanding their great success in image segmentation, F-CNNs require thousands of labeled images for training and their performance degrades when only a small number of annotated images are available [14]. Consequently, an improved mechanism is required for F-CNN training that enables the segmentation of a new semantic class based on a limited number of labeled images [15]. Such approaches frequently use

transfer learning (TL) to transfer the knowledge from pre-trained models to offer an initialization that is later enhanced with the new data to adapt to the underlying problem. Yet, the pre-trained model improvement is still subject to the overfitting problem and requires a reasonable number of labeled images (at least in the order of hundreds) [16]. In situations where there is very little data (such as with COVID-19), the new class has a limited number of annotated images, so such enhancement based TL usually result in overfitting [17]. Few-shot learning (FSL) is an artificial intelligence technique that effectively enables the model to generalize to an anonymous semantic class with a few instances. The primary notion of few-shot learning is driven by an aspect of human learning in which rapid learning of new semantics is possible from a few remarks, exploiting the knowledge acquired from prior experience. Few-shot learning has been extensively studied for object detection and image classification, and lately, used for medical image segmentation. It is shown to be an extremely challenging task to perform pixel-wise predictions in such an extremely low-data regime since it conducts learning from rarely labeled instances since medical experts are required to label images [43] manually. In this paper, we introduce a novel semi-supervised few-shot segmentation (FSS) approach designed specifically for segmenting volumetric COVID-19 CT scans. The key to attain this objective is the combination of the recently proposed recombination and recalibration module within the construction of the proposed architecture.

1.1 Few-shot segmentation

FSL techniques for image segmentation aim to generalize a model to a new observed image with limited annotation using the learned knowledge from various annotated images. The FSL network architecture for image segmentation usually comprises three portions: the conditioner path, segmentation path and the interaction blocks. Throughout the inference procedure, the model is supplied with a pair of images called the support set $(I_s, L_s(\sigma))$ that contains a group of images I_s belonging to the semantic class coupled with their corresponding mask called $L_s(\sigma)$. Simultaneously, a group of unlabeled query images I_q is passed into the model to be segmented. Specifically, the support set is forward fed into the conditioner to produce several feature maps within the middle layers of the conditioner path. These maps are declared as the knowledge representation since they encompass the critical information essential to performing segmentation. The generated knowledge representation is captured via interaction blocks, primarily responsible for passing the pertinent information to the corresponding layers within the segmentation path. Meanwhile, the input query image I_q is fed into the segmentation path that makes use of the transferred knowledge information to produce a segmentation mask M_q. Consequently, the major role of interaction blocks is to communicate the learned knowledge from the conditioner to the segmented and build a powerful architecture for semantic image segmentation. However, most of the present methods [49-51] utilize weak interactions between paths, such as one interaction module at the end layer of the network [16-17].

1.2 Semantic segmentation

Swift advances in medical imaging equipment such as scanners necessitate efficient lesion segmentation techniques that are capable of segmenting the entire infection region and discriminate between relevant interior diseased lesions. Specifically, this necessitates that segmentation approaches must be able to learn more thorough features of various types of infection lesions, which are usually only minor portions of CT images, having an irregular appearance and comparable concentration as the normal areas.

Even though a variety of DL models have offered good solutions for automated lesion segmentation, their lesion segmentation performance requires two crucial enhancements: 1) expanding the receptive field to learn extra features by stacking several convolutions and pooling operations must not result in a decrease in the resolution of feature maps layer-by-layer and hence result in the loss of fractional and small features of the lesion; and (2) owing to the diverse sizes of COVID-19 lesions in the CT images, the DL technique must be able to segment lesions at a range of scales.

In order to address these issues, dilated convolutions (atrous convolution) has been recently employed for capturing multi-scale information in segmentation networks using atrous spatial pyramid pooling (ASPP) [46]. This primarily has two motives. First, it is eminent that the atrous-convolution sample the incoming input data immutably to calculate the output feature map. Second, nearby contexts could be a beneficial type of supplementary statistics to differentiate diverse tissues, counting both the infected and uninfected regions. Despite the efficiency of atrous convolution in capturing multi-scale semantic representation, using it into a segmentation model has two drawbacks degrading the segmentation performance [62,63]: 1) local information loss, since its kernel just performs partial sampling on the nine points of pixels and neglects the pixel values at the in-between sites; and 2) the gridding artifacts problem [39]

1.3 Challenges and Goals

The current computer vision literature for few-shot segmentation (FSS) employs TL from the pre-trained models in both paths to effectively segment RGB images [18]. TL models enable effective exploitation of preceding knowledge with informative features from the beginning of training. Accordingly, adopting a weak interaction module between the conditioner and the segmentation path (i.e., at earlier or later layer) is adequate to train the model efficiently. However, extending such a learning technique to medical images did not realize satisfactory performance due to the absence of the DL model pre-trained on medical data. This limits performance gains realized by TL in medical domains. Hence, we introduce a robust interaction module that enables knowledge communication at several intermediate locations between the

paths while simplifying the gradient flow through the two paths. In view of this, we propose the RR module to communicate the learned representation between the two paths of the FSS network. The module particularly receives the extracted conditioner feature maps as input and performs concurrent spatial and channel squeezing to learn the feature maps from the conditioner path. This is used to accomplish excitation on the corresponding feature map of the segmentation path.

A mutual shortcoming of the most U-Net alike networks is that the strided-convolutions and successive pooling layers gradually decrease the representational resolution to learn the compressed feature representations. Although this behavior is valuable for object detection or classification procedures, it always hinders the segmentation task that necessitates comprehensive spatial representation. Instinctively, maintaining high-resolution feature maps at the intermediate phases can enhance the performance of the segmentation model. Nevertheless, it raises the dimension of feature maps, which is infeasible for accelerating the training operation and facilitate the optimization process. Hence, there is a trade-off between the high resolution and the training speed. In general, the U-Net is shaped with an ED structure. Where the encoder seeks to minimize the spatial size of feature maps progressively and acquire extra complex semantic representations. The decoder seeks to retrieve both the details of the segmentation target (i.e., lesion) and the spatial size. Thus, it is necessary to learn more advanced representations in the encoder and maintain more spatial representation in the decoder to ensure optimal segmentation performance.

Inspired by the debates mentioned above and the Inception-Net [54], the network gets deeper and wider, we propose a novel smoothed atrous convolution (SAC) module. Unlike traditional U-Net architectures that are limited in learning multi-scale representations via 3×3 convolutions and pooling layers through the encoding processes, the proposed SAC can learn and extract a deeper and wider range of semantic representations using four parallel paths of multi-scale smoothed atrous convolutions, while the residual links are employed to avoid gradient vanishing issues. Additionally, we introduce a multi-scale pyramid pooling (MPP) module stimulated by spatial pyramid pooling [55]. The MPP module further learns multi-scale contextual representations of the SAC module entity by employing pooling layers with varied sizes, without requiring any additional learning parameters. Integrating these two modules in the middle of the E-D architecture can help gain greater improvement and reserve extra spatial representation to enhance segmentation performance. The generated E-D architecture is used to build the segmentation and conditioner path.

Furthermore, to avoid the overfitting problem and gain better generalization, we train our model using SSL by incorporating unlabeled CT slices during training. Although most current studies on FSS concentrate on volumetric images with multiple annotated slices, we focus on axial scans of COVID-19. It is time-consuming to manually annotate the lung nodules or infection regions on all slices of CT images of COVID-19 patients. Therefore, we introduce a novel technique, called FSS-2019-nCov, which is able to accurately pair a limited number of COVID-19 slices of the support slices with all the slices of the query set.

1.4 Contributions

The primary contributions of our paper are:

- A novel COVID-19 segmentation technique is based on FSS to enable better generalization from a small number of annotated CT slices in both binary and multi-class scenario.
- We introduce a SAC block and MPP block for efficient exploitation of high-level contextual and spatial information and to assist in overcoming the problem of infection size variation.
- Both the SAC block and MPP block are integrated within the encoder-decoder architecture that is adapted to form the conditioner and segmentation path.
- Adaptive feature recombination and recalibration (RR) modules are included to effectuate knowledge representation interaction between the two paths.
- There is a resultant increase in generalization performance using semi-supervised training for the proposed FSS-2019-nCov.

1.5 Paper Organization

The remainder of the paper is structured as follows. Section 2 reviews the current related studies. Detailed explanations and information corresponding to our proposed frameworks and principles incorporated are presented in Section 3. Proposed experimental conditions, comparison studies, and comprehensive analysis are provided in Section 4. Finally, the conclusions and intended future directions are explained in Section 5.

II. RELATED WORK

In this section, three kinds of studies related to our work are discussed: chest CT segmentation, semi-supervised learning, and COVID-19 segmentation.

A. Chest CT segmentation

The CT scan is a prevalent diagnostic tool for lung diseases [6]. Practically, segmenting a variety of lesions from chest CT images supplies clinicians with substantial information on lung disease diagnosis and quantification [19]. Several studies achieve chest nodules segmentation using a feature extractor accompanied by a classifier. For instance, Kumar et al. [20] introduced a new supervised CNN to fuse complementary multi-modality information from lung cancer scans. Ozdemir et al. [21] addressed lung cancer diagnosis using a 3D probabilistic DL approach for nodule segmentation and diagnosis while presenting model uncertainty. Gerard et al. [22] proposed a coarse-to-fine cascade of two CNN to reduce the impact of thin structure on the segmentation network. Jiang et al. [23] developed two residual networks to concurrently syndicate features across several resolutions and levels to detect lung tumors. Additionally, Cheplygina et al. [24] reviewed the

recent studies of semi-supervised techniques for medical imaging tasks. However, these studies are extremely successful in data-intensive problems but are often obstructed for very small data sets. Such approaches also suffer from low generalization capability, making them inefficient for the underlying task of COVID-19 segmentation. To tackle these issues, we propose an FSL-based approach to enable learning from limited data.

B. Few-Shot Segmentation (FSS)

Recently, many studies have explored FSS with deep learning. Caelles et al. [49] performed video segmentation using the first frame annotation based on the notion of tuning pre-trained architectures. Even though their model operates effectively in this scenario, it is subject to overfitting and necessitates retraining to adopt a new class, which hampers the swiftness of adaptation. Shaban et al. [16] introduced a two-step approach, where the first step processes the new image-label pair to infer the classification parameter for the other step, which receives a query image and predicts the corresponding segmentation. Dong et al. [50] improved this approach to address numerous unidentified classes to perform multi-class segmentation simultaneously. Rakelly et al. [51] applied the approach in a very difficult scenario and they chose a tiny set of landmarks to induce the supervision of the support set, rather than using a compactly annotated binary mask. The training process in the before-mentioned studies relies on TL models. Despite the effectiveness of TL in many computer vision studies, there are no pre-trained architectures in medical imaging.

In the medical imaging domain, FSS was first proposed in [52]. The authors used adversarial learning for brain image segmentation depending on one or two annotated labeled brain images, enthused by the previous achievement of SSL. Zhao et al. [53] exploited the captured transformation to extremely augment a fully labeled volume for one-shot segmentation. Roy et al. [43] introduced the two-stage model and applied the recently proposed squeeze and excite modules to empower the knowledge exchange between both arms and smooth the gradient flow. However, these studies suffer from a number of shortcomings. First, the approaches in these papers rely on the assumption that every shot is a complete 3D image that comprises many 2D slices. Second, they construct huge architectures without analyzing the effectiveness of every building block, which results in composite and potentially inconsistent models. Finally, they considered neither contextual information nor multi-scale features.

Motivated by this, our study investigates the role of unsupervised data in the process of segmenting COVID-19 CT scans in an FSL scenario. Predominantly, we make use of the successful achievement of FSS studies in normal images. To further boost the performance of the proposed FSS-2019-nCov, we leverage unannotated axial CT slices as a supervisory signal. Incorporating unannotated slices into auxiliary tasks has been used to improve the generalization capabilities of deep learning approaches in many studies.

C. Semi-supervised Learning

While owing to the challenge in finding entirely annotated data, semi-supervised learning (SSL) has been attracting much attention to enhance the network performance using a small amount of annotated data and a very large amount of unlabeled data [24]. SSL has been widely adopted for training deep models, which always seek to optimize the supervised and unsupervised loss on labeled and unlabeled data, respectively. [25-26]. Fan et al. [27] proposed using weighted Intersection-over-Union (IoU) loss for edge supervision and cross-entropy loss for segmentation supervision. In a nutshell, the current deep SSL models regulate the network by imposing fine-grained and reliable classification boundaries, which are vigorous to a random disturbance. Other approaches enhance the supervision signals by investigating the acquired knowledge, such as pseudo labels and temporal ensemble dependency [28]. Inspired by the recent success of SSL architectures in the studies mentioned above, we propose to adopt SSL in model training to attain better generalization and avoid overfitting effects that may be incurred with pre-trained models.

D. COVID-19 Segmentation

Recently, artificial intelligence has been widely adopted in multiple applications applied to COVID-19 detection [29]. These applications could be categorized into three groups [30]: patient-level, concerned with medical images analysis tasks (e.g., segmentation, classification, and quantification, etc.); molecular applications dedicated to protein structure (e.g., protein interactions, drug repurposing, etc.); and societal applications concentrated on epidemiology related tasks. In this paper, we focus on patient-relevant applications, specifically for those depend on CT scans. For example, Wang et al. [31] introduced an adapted inception network to classify COVID-19 patients from normal cases by training the network on the ROIs defined with two experienced radiologists according to the characteristics of pneumonia. Chen et al. [32] accumulated 46,096 slices of CT volumes from confirmed COVID-19 cases and other disease cases. The collected CT slices were used to train a U-Net++ [12] to identify COVID-19 cases. Their results demonstrate that the model diagnoses COVID-19 as well as radiologists. Additionally, other models proposed to act as AI-assisted systems for COVID-19 diagnosis, including ResNet [33-34], and U-Net [27][31]. Moreover, deep learning has been utilized for infection segmentation in lung CT scans and the obtained quantitative outcomes can be exploited to assess disease severity [35], quantify infection [3], screen infection at a large-scale [36]. All of the studies mentioned above assumed utilizing a large amount of data to train their models in a supervised manner, but the lack of annotated CT scans for COVID-19 means that such approaches lack utility. Fan et al. [27] were first to tackle this problem using a semi-supervised learning scheme, yet they first segment infection regions to use them to guide the multi-class segmentation, which results in suboptimal performance. This motivates us to use FSS to enable better generalization from small data samples using newly proposed context encoder-decoder architecture, efficiently exchanging this knowledge with segmentation path using the proposed smoothed RR module. We also boost the

model generalization by semi-supervised training incorporating unlabeled CT slice.

III. PROPOSED APPROACH

In this section, we present a detailed explanation of the proposed FSS-2019-nCov in terms of network architecture, network building blocks, and cost function. Then, we introduce the semi-supervised variant of FSS-2019-nCov and clarify how to use an SSL paradigm to increase the number of training instances to improve the segmentation performance. In addition, we extend our model for the multi-label classification for a variety of lung infections. Finally, we indicate the details of the implementation.

A. Problem Formulation for FSS

In the infection segmentation scenario, the training data for FSS-2019-nCov $D_{Train} = \{(I_T^i, L_T^i(\propto))\}_{i=1}^{N}$ contains N couples of CT axial scan and its respective annotation map $L_T(\propto)$. In the multi-class scenario, every semantic label $\propto \in L_{Train}$ have an annotation map $L_T^i(\propto) \in D_{Train}$ where $L_{Train} = \{1,2,\ldots,K\}$, where k is the number of training classes. (e.g., in multi-class COVID-19 segmentation, the 1, 2, and 3 represent the GGO, consolidation, and background). The FSS-2019-nCov learns on D_{Train} with objective function $F(\cdot)$ such that having a support set $(I_s, L_s(\hat{\propto})) \notin D_{Train}$ for a new semantic label $\hat{\propto} \in L_{test}$ (L_{test} is the number of testing class) and a query slice I_q, the COVID-19 infection segmentation $M_q(\hat{\propto})$ of the query is predicted. There is no intersection between semantic labels of training and testing data. The most remarkable aspect of FSS is that test classes L_{test} exist in the training data as the background class. This possibly could be exploited as a form of past knowledge during testing in cases where scarce instances are provided with the infection annotated.

B. FSS-2019-nCov Architecture

As previously stated, the architecture of FSS-2019-nCov comprises three modules: the conditioner path, the adaptive interaction module, and the segmentation path. The conditioner path learns the visual information of the support set to infer infection on the query slice. The adaptive interaction module effectively conveys the learned representation in terms of feature maps to the segmentation path. The segmentation path makes use of the acquired representation to segment the query slice. Figure 1 shows the detailed architecture of the proposed FSS-2019-nCov, which is further described in the following subsections. In FSS-2019-nCov, both the conditioner and the segmentation have an identical layout. In this way, feature maps in each path have the symmetric spatial resolution, which facilitates and empowers the interaction between corresponding blocks and eases gradient flow.

1) E-D Architecture of Conditioner and Segmentation Paths

The architecture of the conditioner path has an encoder-decoder based architecture consisting of four encoder blocks based on pre-trained Res2Net [37], four decoder blocks, and a Context enrichment (CE) module —see Figure. 1.

Feature encoding: Recently, the Res2Net [37] architecture has shown great success in many computer vision tasks and has validated its effectiveness overall residual architectures owing to the multi-scale feature extraction capability that enables fine-grained level representations for every network layer. Motivated by this, we propose to implement the encoder path using Res2Net-50 as a backbone architecture. The structure of the encoder (or Res2Net) module is presented in Figure 2 (a).

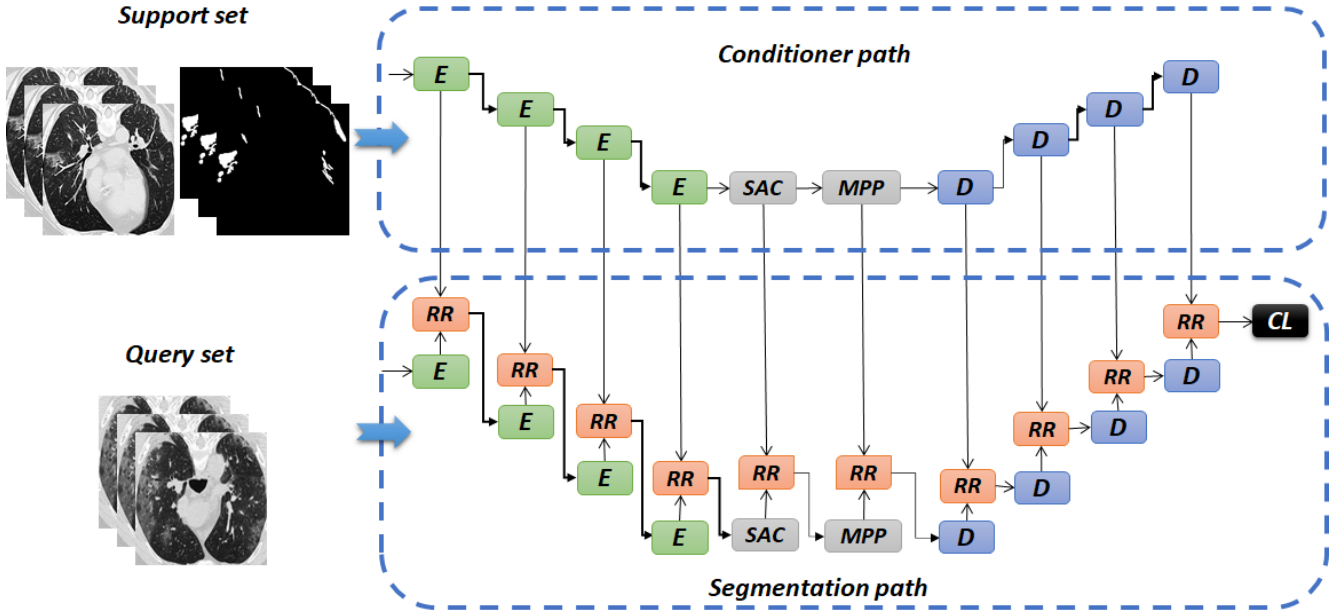

Figure 1. The architecture of the proposed FSS-2019-nCov. It consists of two identical paths with the encoder-decoder structure, namely the conditioner path (upper) and the segmentation path (top). The recombination and recalibration (RR) blocks (see Figure 4) are introduced to effectuate knowledge interaction between two paths. The axial CT images are passed through a feature encoder blocks (E) module that is implemented with the pre-trained ResNet-34 blocks. The context enrichment module is then introduced to generate an improved semantic representation using SAC and MPP modules. Finally, the acquired representations passed into the feature decoder blocks (D).

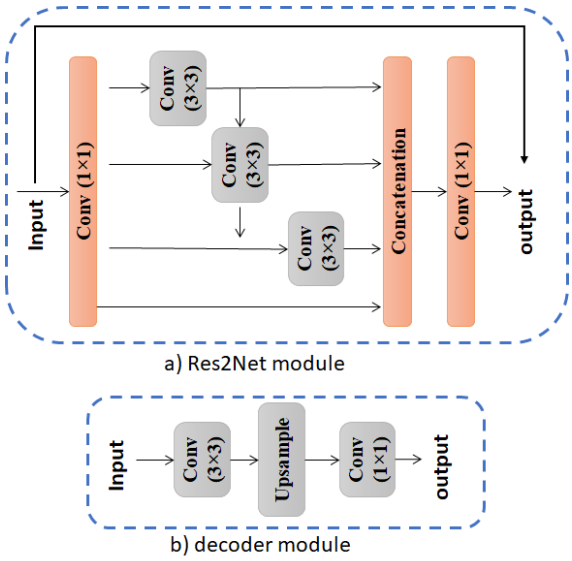

Figure 2. Illustration of the encoder and decoder modules used in the proposed FSS-2019-nCov: a) the encoder module implemented using Res2Net module [37]; and b) the architecture of the decoder module

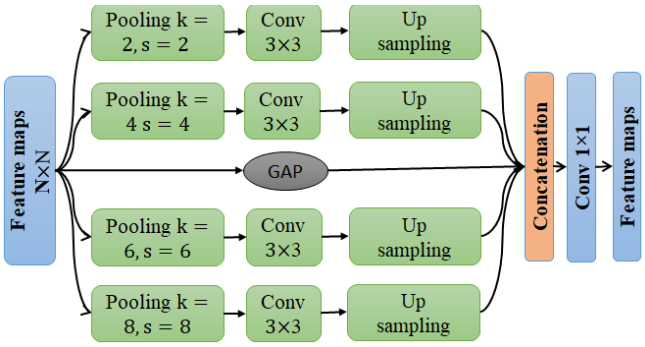

Figure 4. The architecture of the proposed MPP module containing five parallel paths for changing input resolution. Convolution layers are employed to capture different resolution information. The global average pooling (GAP) layer is employed to implement the residual connection.

the multi-scale processing enables learning more representative information from the input CT images. The residual linking facilitates the network convergence and evade the gradient vanishing problem. From the input image, the E blocks acquire the global representation of the target entity (i.e., lesions) and relevant parts class property of the target [26], [32]. Nevertheless, these kinds of representation might gradually debilitate at the time they transmitted to deeper levels [24]. Thus, we introduce the CE module to tackle this issue, as presented in Figure. 1.

Context Enrichment: The CE is introduced to learn semantic context representation and hence provide more informative feature maps, and it contains two blocks: the smoothed atrous convolution (SAC) block and the multi-scale pyramid pooling (MPP) block.

Smoothed Atrous convolution: The typical convolutional layer is widely adopted feature extraction in many semantic segmentation tasks [46]. Nevertheless, it still suffers from semantic information loss caused by pooling layers. In order to tackle this shortcoming, atrous convolution has been used in

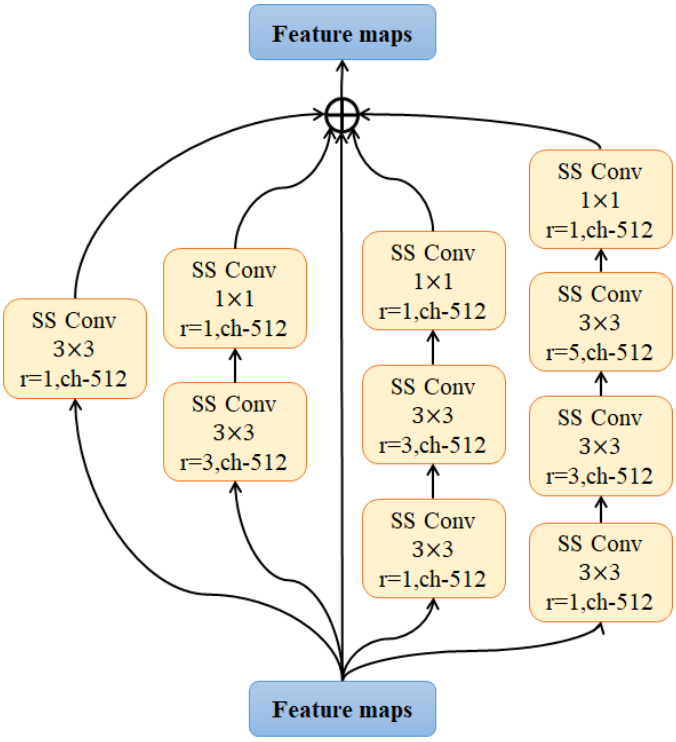

Figure 3. The architecture of the proposed SAC module consisting of four parallel paths. Each path from left to right contains 1,2,3, and 4 separable and shared convolutions, respectively.

many segmentation tasks [38]. However, atrous convolution (with dilation larger than 1) still suffering from the gridding artifacts problem [39], which means that the calculation of neighboring is based on dispersed sets of input units, which causes local information discrepancy and degrades the network performance. This issue has been tackled with the recently proposed separable and shared convolution (SS-Conv) [39]. In an attempt to capture multi-level information learned through an encoder, we propose the SAC block presented in Figure. 3. in which we stack the SS-Conv layers in the form of four cascade tracks with a receptive field of 3, 7, 9, 19 causes a gradual increase in the number of SS-Conv from 1 to 1, 3, and 5. Inspired by the inception module [47], we attach SS-Conv 1×1 with *Relu* activation at the end of each track. Finally, we concatenate the output of four tracks with the original feature maps as the output of the SAC block. The SS-Conv with a large reception field effectively captures and produces more detailed information for large infection areas. In contrast, the SS-Conv with a small reception field is better for small infection areas. By integrating the atrous SS-Conv of various atrous rates, the SAC enables efficient feature extraction for infections of various sizes.

Multi-scale pyramid pooling: The most challenging issue in infection segmentation is the wide variety of infection sizes in medical scans. For instance, the size of GGO in the middle or late stage can be much larger than that in the early stage of COVID-19 infection [5-6]. To tackle this problem, we propose multi-scale pooling layers that depend on several operative fields of view to distinguish infection of various sizes, as shown

in Figure 4. Unlike [55], MPP takes the incoming feature maps and passes themto the four paths to alert their resolution using the pooling layer (i.e. average or max), hence the resolution at each path gets decreased to 1/2, 1/4, or 1/8 of the corresponding input. Then, a 3×3 convolution is employed to extract and learn multi-scale contextual representations. Additionally, we redesign the residual connection [55] to be implemented with global average pooling (GAP). Unlike [55], the MPP module can capture extra contextual information from the received input due to the nature of the average pooling operation that processes input maps at the regional level instead of point level. For example, given an input map 64×64, decreasing the input map resolution to 1/8 creates the new map of 16 (*i.e.*, 4×4) points, so that 3×3 convolutions could capture information of nine of them, which means increasing the information consumption ratio. Thus, applying such a pooling layer enables the utmost input map values to contribute to the output map of the MPP module. Additionally, reducing the input map resolution often decreases the computation burden and logically increases the time efficiency compared with [55]. Further, the proposed non-dilated convolution usage in the MPP module also helps avoid the gridding artifacts problem [39]. After the convolution layer, the low-dimension feature map is up-sampled using bilinear interpolation to obtain the feature map with the same size as the input feature map. Furthermore, similar to SAC, the up-sampled feature maps are concatenated with the input feature map. Finally, the concatenated map is passed into then 1×1 convolution to generate the final output of the MPP module. Optimal parameter grid search showed that the size of stride should be to 2, 4, 6, and 8, which corresponds to kernel dimensions of 2, 4, 6, and 8, respectively.

Feature Decoding: For restoring powerful resolution feature representations rapidly and professionally, four simple D blocks are employed to form the decoder path. The main purpose of the decoder is to reinstate the spatial representation with sophisticated features engendered fromthe CE module and progressively fuses the global contextual information. The architecture of decoder blocks presented in Figure 2 (b) contains 3×3 de-convolution, followed by a sampling layer for reducing the number of network parameters. The output of a D block is attained after 1×1 convolution. The generated map of the last D block is directly up-sampled to the same dimension of the original image. Therefore, the D blocks have the following number of filters 64, 128, 256, and 512 sequentially.

2) Conditioner Path

The main job of ask of the conditioner path is to take as an input the support set with a slice I_s and mask L_s, which is later passed to the proposed encoder-decoder architecture to learn the visual representation that is used to generate informative task-specific feature maps and enable detecting the area to be segmented in the query slice I_q in the segmentation path. In this paper, the feature maps of the middle layers of the conditioner path are referred to as the knowledge representation. The conditioner path has a two-channel input formed by stacking I_s and L_s ($\propto$).

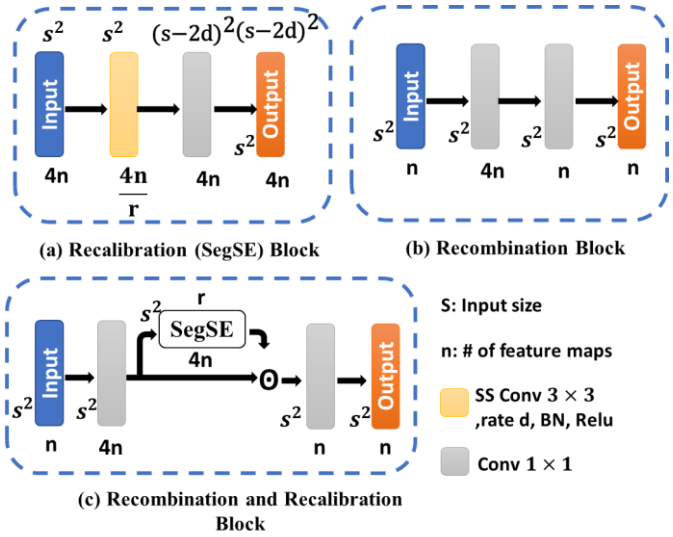

Figure 5. The architecture of the RR module: a) illustration of the recalibration block implemented using separable and sharable convolution; b) illustration of the recombination block; and c) integration of both recalibration and recombination in a single module.

3) Adaptive Interaction Module

The interaction module plays an essential role in the proposed for FSS-2019-nCov. It consists of multiple interaction blocks that take the generated knowledge representation of the conditioner path as input and transfer it to the segmentation path to conduct the query slice segmentation. The most essential characteristics of these blocks are 1) a slight increase in the computational complexity of the model; 2) improved gradient flow and hence facilitated model training, and 3) adaptive exploitation of channel-wise relationships. For this purpose, we introduce a modified version of the recently proposed feature recalibration block (SegSE) and combine it with the feature recombination block [40] to obtain the recalibration and recombination (RR) module presented in Figure. 5 (c). SegSE blocks are computational blocks to achieve adaptive recalibration of feature maps that act as a channel-wise attention mechanism that improves the discriminative power of generated feature maps, with a marginal increase in model complexity.

Recalibration Module: Since there is a spatial correspondence between the segmentation pixels/voxels and the units of feature maps, applying channel squeeze and excite (SE) operation [41] potentially suppresses the entire feature maps that could encompass significant regions. To address this, we propose to use a spatially adaptive variant of SE (SegSE) that enables concurrent spatial and channel SE, which is more appropriate for COVID-19 semantic segmentation. The architecture of the SegSE block is presented in Figure. 5(a). The spatial structure and the correspondence of the feature maps are preserved by replacing the global average pooling in the SE block with $SS\ Conv\ (3 \times 3)$ layer to capture large-scale contextual information through dilated kernel operating over adjacent voxels to obtain Z^{SegSE}, but without increasing those kernels' parameters. Assuming that the convolution layer performs the transformation function F that maps the input X to the output U where $X \epsilon \mathbb{R}^{H' \times W' \times C'}$, $U \epsilon \mathbb{R}^{H \times W \times C}$; H', W'

represents the height and width of the input feature map; H, W represents the height and width of the output feature map and C, C' denote the count of feature maps such that $X = \{x_1, \cdots x_{C'}\}$ and $U = \{u_1, \cdots u_C\}$. Then, we obtain a feature map $\boldsymbol{Z^{SegSE}}$ using equation (1-2).

$$Z^{SegSE} = \gamma\left(F^{conv}(X; k^{segSE}, d, n^{segSE})\right) \quad (1)$$

$$n^{segSE} = \frac{C'}{r} \quad (2)$$

where γ denotes the batch-normalization tailed with the ReLU activation function, k is the kernel size, d represents the dilation rate that is determined based on the scale of the layer, n is the number of kernels, and r denotes the reduction factor. Hence, increasing the number of $conv$ layers increases the field of view, which means that the units of the feature maps represent a wider area of the input space. After that, to obtain the recalibration output feature maps, a convolutional layer with kernels 1×1 operates on Z^{SegSE}, and its output is fed into the sigmoid function as formulated in equation (3).

$$S = \sigma\left(F^{conv}(Z^{SegSE}; k, d, n)\right) \quad (3)$$

where $k = 1, d = 1$, and $n = C'$. Thus, we integrate the squeeze and excitation operation since the dilated $conv$ layer decreases the number of feature maps, presenting a bottleneck. Finally, element-wise multiplication Θ is applied to input S to obtain the recalibrated feature maps. So, the recalibration of the given feature map c is calculated with equation (4).

$$u_c = x_c \odot s_c \quad (4)$$

So, the overall operation of the SegSE block could be expressed as $F^{segSE}: X \rightarrow U^{segSE}$

Recombination Module: The main purpose of recombination is to empower the representativeness of the features by linearly combining them (see Figure.5 (b)). Accordingly, we utilize a convolutional layer with a kernel size of 1×1. The features map F^{exp} is expanded with factor m and then recompressed again to the original number size F^{comp}. Thus, recombination operation could be expressed as $F^{recomb}: X \rightarrow U^{recomb}$ where $U^{recomb} \in \mathbb{R}^{H' \times W' \times C'}$ is mathematically formulated in equation (5).

$$F^{recomb} = F^{comp}(F^{exp}(X, mC'), C') = F^{comp}(F^{exp}(X, 1,1, mC'), 1,1, C') \quad (5)$$

4) Segmentation path

The main target of the segmentation path is to segment the input query slice I_q utilizing the knowledge representation acquired from the conditioner path, which contains a high-level informative feature about the formerly unseen query slice. The SegSE blocks within the interaction modules compress the feature maps of intermediate layers of conditioner. They then perform cross-channel feature recalibration on the feature maps of the segmentation path. The architecture of the segmentation path is symmetric to the conditioner with just two main variations: 1) unlike the segmentation path, there are no interaction blocks presented after encoding and decoding modules of the conditioner path.; and 2) in the segmentation path, we final classification block with $Conv$ (1×1) layer that produces the output segmentation maps that followingly fed into $Softmax$ function to infer the infection segmentation in query slice.

C. Semi-supervised training

Currently, there are only a small amount of annotated CT images for COVID-19 patients. The manual segmentation of lung area and COVID-19 lesions is laborious and time-consuming, and most studies focus on studying the virus itself and finding the best inhibitor. To tackle this data limitation problem, we propose to train the FSS-nCoV-Net in a semi-supervised manner, in which the widely available unannotated CT image set is exploited for augmenting the training data, motivated by recent studies in [60,61,27], in which a random sampling mechanism for gradually expanding the CT training data using unannotated CT images. Algorithm 1 is employed unambiguously to estimate and generate the pseudo labels corresponding to the unannotated CT images. The follow-on CT scans, along with the corresponding pseudo labels, are subsequently exploited to train the proposed FSS-2019-nCov. In view of this, semi-supervised training of the proposed FSS-201-nCov has several benefits summarized as follows. First, the training and assortment technique is straightforward and not difficult to implement. Second, it is threshold-free and also does not necessitate measures to evaluate the forecast annotation. Third, it helps avoid the overfitting issue, which can provide more robust performance than other semi-supervised training approaches demonstrated by recently published studies [60, 61, 27].

Algorithm 1 Semi-supervised training for the FSS-2019-nCoV

Input: 1) $D^{\text{Annotated}}$ represent the annotated CT slice in the COVID-19 dataset; 2) $D^{\text{Unannotated}}$ represents the unannotated CT images in the COVID-19 dataset.

Output: Trained FSS-2019nCov

1: Split the data to create the training dataset D^{Training} utilizing all the annotated CT images from $D^{\text{Annotated}}$.
2: Start training the proposed FSS-2019-nCov denoted as F using D^{Training}.
3: **While** $D^{\text{Unannotated}}$ is not empty, **do:**
4: Test the trained model F on N CT slice arbitrarily nominated from $D^{\text{Unannotated}}$.
5: Generate model-annotated data $D^{\text{F-annotated}}$ containing N CT slice with pseudo labels
6: Use $D^{\text{F-annotated}}$ to expand the training set, such as $D^{\text{Training}} = D^{\text{Training}} \cup D^{\text{F-annotated}}$
7: Eliminate the N testing CT slice from the $D^{\text{Unannotated}}$
8: Fine-tune the model F *on the* D^{Training}.
9: **End While**
10: **Return** Trained Model F

D. Model training methodology

We train our model using the training mechanism adopted in [16,43], where a batch sampler is used to randomly select a mini-batch that is subsequently used for model training. As opposed to traditional supervised training, we implement the following steps for picking samples from a mini-batch in every iteration. First, a label $\propto \in L_{Train}$ is randomly selected. Second, two CT slice and their corresponding labels are randomly sampled, such that they contain a semantic label $\propto$. Third, binarization of the label map to set label $\propto$ at the foreground and to make the remaining areas the background. Fourth, the two pairs respectively establish the support set $(I_s, L_s(\propto))$ and the query set $(I_q, L_q(\propto))$, where $L_q(\propto)$ is the GT for calculating the loss. To sum up, the FSS-2019-nCov takes the two pairs as a training batch, where the support pair $(I_s, L_s(\propto))$ is combined to form two-channeled input to the conditioner path. Meanwhile, the query slice I_q is used as the segmentation path input. Both inputs pass through the two paths of the model in a feed-forward manner seeking to predict the segmentation $M_q(\propto)$ for the query slice I_q for label $\propto$. Dice loss [44] calculated between $M_q(\propto)$ and $L_q(\propto)$ using equation (6) is:

$$L_{Dice} = 1 - \frac{2\sum_x M_q(\propto) L_q(\propto)}{\sum_x M_q(\propto) + \sum_x L_q(\propto)} \quad (6)$$

where x represents the pixels of the prediction map. In order to reduce the L_{Dice}, the batch sampler offers different instances belonging to diverse $\propto$, and the loss is calculated for that particular $\propto$ and subsequently, the weights are modified, continuous altering of the inputs at each iteration, makes the model converges. Therefore, it could be said that the prediction turns out to be agnostic to the selected $\propto$.

We train FSS-2019 to minimize the L_{Dice} loss for segmentation from annotated slices only. Simultaneously, to leverage the unannotated CT slices data, we employ an auxiliary manifold embedding loss L_E on the dormant feature representations $h(\cdot)$ of both labeled and unlabeled samples to diminish the discrepancy between similar inputs in the latent space [45]. Thus, similarity among $h(\cdot)$ of unlabeled CT slices is specified by preceding knowledge. The final objective function could be formulated using Lagrangian multipliers, as shown in equation (7).

$$L_{total} = L_{Dice} + \sum_x R_l \cdot L_{E_l} \quad (7)$$

where R_l represent regularization parameter for the embedding loss E_l at hidden layer l. Naturally, this loss function seeks to minimize the distance between concealed representations of analogous $h^l(x_i)$ and $h^l(x_j)$ of adjacent data samples and, if not, attempt to push them away from each other. Furthermore, through extensive experiments, we tried different model training parameters to find out the most optimal configuration for our model and got the highest performance using the parameter shown in Table 1.

Table 1. model training parameters

Methods	DSC
Learning Rate	0.001
Weight decay constant	0.0001
Momentum	0.9
No. of epochs	50
Iterations per epoch	300
Optimizer	SGD
Balance factor	0.5

IV. EXPERIMENTS AND RESULTS

A. Dataset

Two annotated CT datasets are employed for model evaluation, publicly published by the Italian Society of Medical and Interventional Radiology [42]. The first dataset (CT-1) comprises 110 axial CT slices belonging to 60 patients that are positively confirmed to have Covid-19. The CT slices were greyscaled, resized, and compiled into a NIFTI-file. The size of each slice was set to 512×512 pixels. An experienced radiologist annotated the CT slices using three-class labels, namely pleural effusion, GGO, and consolidation. We eliminated two images because of their low resolution. We split the CT-1 data into a training set of 38 CT images, a validation set of 20 images, and a test set of 50 images. Additionally, the second dataset (CT-2) comprised nine CT volumes consisting of 829 slices. Among them, there were 373 annotated axial CT slices that were positively confirmed as a COVID-19. 638 axial slices (i.e. 285 lesion-free slices and 353 infected slices) were selected for model evaluation. The annotated CT slice was resized from 630×630 resolution to 512×512 resolution as with CT-1 data. For semi-supervised training, a total of 1600 unannotated axial CT images were collected from the COVID-19 CT dataset [59], comprising 20 CT volumes from distinct COVID-19 patients. Then, the data was prepared by eliminating non-lung regions to form an unlabeled training set. All slices were preprocessed using an intensity normalization procedure for all input data.

B. Comparative Studies

Baseline architectures. In the experiment relevant infection region segmentation scenario, we compare our model with robust semantic segmentation models including UNet [9][1] and H-DenseUNet [11][2], U-Net++[12][3], SegNet [13][4], FCN8s [7][5], DeepLabV3+[14][6], SE-Net[43][7] Inf-Net [27] as a baseline architecture, and compare the proposed approach against the recently proposed Inf-Net for COVID-19 segmentation [27]. In the multi-class scenario, we compare the proposed FSS-2019-nCov against the before mentioned, including DepLabV3+ [14] with different stride values, FCN8s [7], and Semi-Inf-Net-U-Net [9], Semi-Inf-Net-FCN8s [27], Semi-Inf-Net, and MC[27].

[1] https://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/
[2] https://github.com/xmengli999/H-DenseUNet
[3] https://github.com/MrGiovanni/UNetPlusPlus
[4] https://github.com/alexgkendall/caffe-segnet
[5] https://github.com/BVLC/caffe/wiki/Model-Zoo#fcn
[6] https://github.com/tensorflow/models/tree/master/research/deeplab
[7] https://github.com/abhi4ssj/few-shot-segmentation.

Table 2. Model comparison for COVID-19 infection segmentation

Methods	Pre-trained architecture	DSC↑	Sens↑	Spec↑	S_α ↑	E_ϕ ↑	MAE↓
U-Net [9]	VGG16	0.459	0.568	0.881	0.639	0.651	0.196
H-DenseUNet [11]	DenseNet-101	0.537	0.611	0.870	0.663	0.683	0.189
U-Net++ [12]	VGG16	0.607	0.701	0.932	0.739	0.751	0.139
SegNet [13]	VGG16	0.657	0.728	0.941	0.744	0.750	0.129
Inf-Net [27]	Res2Net	0.705	0.746	0.966	0.798	0.851	0.086
SE-Net [43]	-	0.621	0.719	0.949	0.751	0.801	0.142
Semi-Inf-Net [27]	Res2Net	0.752	0.757	0.965	0.818	0.902	**0.061**
***FSS-2019-nCov**	Res2Net	**0.798**	**0.803**	**0.986**	**0.834**	0.908	0.065

↑ denote 'higher is better', ↓ denote 'lower is better'

C. Evaluation Metrics.

In this study, we choose three broadly adopted metrics for performance evaluation namely Sensitivity (Sen.) $= TP / (TP + FN)$, *Specificity* (*Spec.*) $= TN / (FP + TN)$ and Dice similarity coefficient (DSC).

In order to measure the overlap between the segmentation outcomes represented with set S and the ground-truth represented with set G, the DSC is calculated as formulated in equation (8).

$$DSC = \frac{2|S \cap G|}{|S||G|} \tag{8}$$

where $|.|$ denote the set size, and $S \cap G$ denotes the intersection of both sets. The generated score always exists between 0 and 1; achieving high DSC reflects the greater segmentation performance.

Also, following [27], we adopt three additional object detection metrics as follows.

1) The structural similarity between a calculated map and the GT mask is measured with *Structure Measure* (S_α) with balance factor α between object-aware resemblance (S_o) and object region-aware resemblance (S_r) according to equation (9).

$$S_\alpha = (1 - \alpha) * S_o(S_p, GT) + \alpha * S_r(S_p, GT) \tag{9}$$

Here, we choose $\alpha = 0.5$, as recommended by the original study [56] and some other recent studies either for COVID-19 segmentation [27], semantic segmentation [57], or object detection [58].

2) The recently proposed *Enhanced-alignment Measure* (E_ϕ^{mean}) to measure similarity (local and global) between two maps based on equation (10).

$$E_\phi = \frac{1}{w \times h} \sum_x^w \sum_y^h \phi(S_p(x,y), GT(x,y)) \tag{10}$$

where w and h respectively represent the width and height of GT, the (x, y) is the pixel position in GT, and ϕ denote the boosted alignment matrix. The value of E_ϕ calculated transforming the prediction S_p into a binary mask with a threshold value in the range [0,255] as introduced in [48]. We provide the average of E_ϕ calculated from overall thresholds.

3) *Mean Absolute Error (MAE):* used to compute the error between S_p and GT at the pixel level as formulated in equation (11).

$$MAE = \frac{1}{w \times h} \sum_x^w \sum_y^h |S_p(x,y), GT(x,y)| \tag{11}$$

D. Results and discussion

1) *Whole lung infection segmentation*

In Table 2, we present the obtained results of the proposed FSS-2019-nCov on the five before-mentioned metrics. It could be observed that our model performs COVID-19 infection segmentation with DSC of 0.789, the sensitivity of 0.803, Specificity of 0.986, $\boldsymbol{S_\alpha}$ of 0.834, $\boldsymbol{E_\phi}$ of 0.908, and MAE of 0.065, which outperforms the cutting-edge studies on the first four metrics. Also, it could be observed that the SSL based architectures (i.e. Inf-Net [27], Semi-Inf-Net [27], and the FSS-2019-nCov) have the highest performance on all metrics compared to the supervised models that require a large number of samples to learn. This supports our choice for training FSS-2019-nCov in a semi-supervised manner. In addition, the FSS-2019-nCov achieved 4%, 5%, 2%, and 2% improvement respectively on DSC, Sens, Spec, and $\boldsymbol{S_\alpha}$ over the recently proposed Semi-Inf-Net, which validates the effectiveness of FSS for tackling problems with low volumes of data. Besides, that Semi-Inf-Net still shows the lowest MAE. This might be explained by the negative impact of eliminating the skip connection in our E-D architecture, which also demonstrates the effectiveness of GT guidance presented in [27].

In addition, we can further confirm the effectiveness of semi-supervised FSS-2019-nCov by providing a visual comparison of the output of different models, as presented in Figure. 6.

2) *Multi-Class scenario*

In addition to whole lung segmentation, we seek to provide more informative segmentation of different classes of lung infections, namely GGO, which is represented as a hazy grey shade, and consolidation is represented as opacification with obscuration of margins. Thus, we evaluate the proposed FSS-2019-nCov in the context of multi-class lung infection to validate the efficiency of the model in providing clinicians with fine-grained information for COVID-19 diagnosis and quantification. Table 3 presents the quantitative results of the multi-class FSS-2019-nCov on GGO class compared with state-of-the-art approaches. For GGO lesion, the FSS-2019-nCov achieved 0.679 of DSC, 0.768 of Sens,0.980 of Spec, 0.735 of $\boldsymbol{S_\alpha}$, 0.894 of $\boldsymbol{E_\phi}$, and 0.061 of MAE. It could be noted that the supervised models (i.e., FCN8 and DeepLab V3+) with pre-

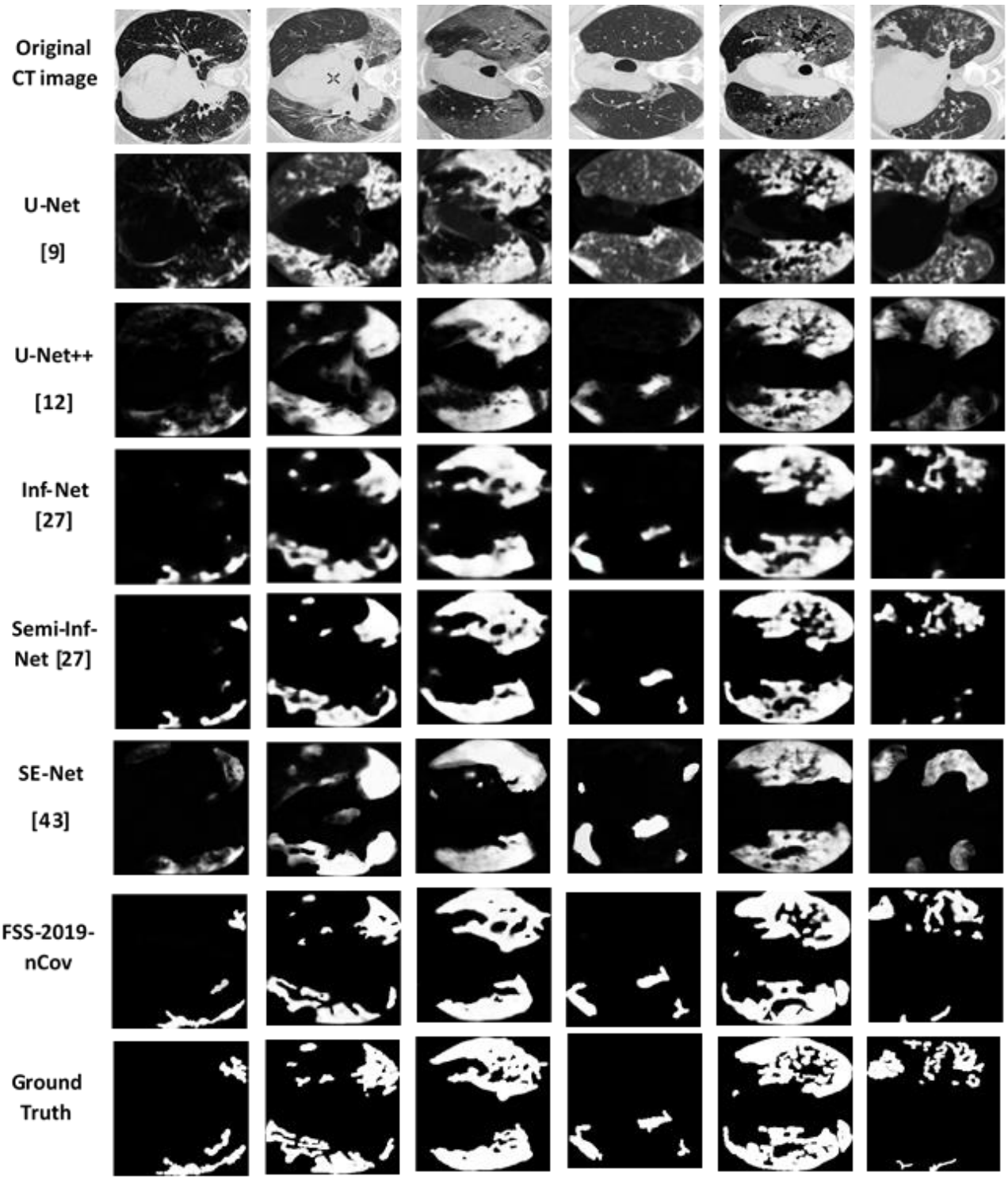

Figure 6. Lung infection segmentation using proposed FSS-2019-nCov. The first row represents the original CT image from the test set. The corresponding segmentation outcome from the U-Netv [9], U-Net++[12], Inf-Net[27], Semi-Inf-Net[27], SE-Net[43] are presented in the second, third, fourth, fifth, sixth row respectively. The segmentation results of the proposed FSS-2019-nCov is presented in the seventh row. The corresponding ground truth label for every image is presented at the bottom of the last row of images.

trained backbones show unacceptable performance owing to the data-hungry nature of supervised learning. Among them, a multi-class version of U-Net [9] shows comparatively higher results on several metrics. Additionally, few-shot-based SE-Net [43] has shown 3% improvements on the DSC measure though in the absence of a pre-trained backbone, which explains the superiority of few-shot learning limited data scenarios. Moreover, the semi-supervised approaches (either Semi-Inf-Net-FCN8s or Semi-Inf-Net MC) shows better performance than supervised models or few-shot based SE-Net [43]. This explains the effect of incorporating unlabeled samples in training to improve model classification performance and improve generalization performance. Furthermore, we also note that FSS-2019-nCov obtains 2.2%, 3.7%, and 1.7%

Table 3. Model comparison for GGO segmentation

Methods	Pre-trained architecture	GGO segmentation						Consolidation segmentation					
		DSC↑	Sens↑	Spec↑	S_α ↑	E_ϕ ↑	MAE↓	DSC↑	Sens↑	Spec↑	S_α ↑	E_ϕ ↑	MAE↓
FCN8s [7]	VGG16	0.482	0.552	0.917	0.591	0.788	0.098	0.289	0.281	0.728	0.573	0.581	0.058
DeepLabV3+ (s=8) [14]	ResNet101	0.402	0.501	0.871	0.553	0.682	0.121	0.157	0.173	0.744	0.511	0.556	0.065
DeepLabV3+ (s=16) [14]	ResNet101	0.457	0.728	0.845	0.559	0.673	0.149	0.245	0.322	0.721	0.526	0.619	0.079
U-Net [9]	VGG16	0.462	0.374	**0.988**	0.564	0.731	0.079	0.421	0.427	0.978	0.581	0.781	0.053
SE-Net [43]	-	0.508	0.415	0.889	0.541	0.751	0.075	0.449	0.467	0.958	0.554	0.797	0.051
Semi-Inf-Net-FCN8s [27]	Res2Net + VGG16	0.657	0.731	0.954	0.722	0.884	0.073	0.318	0.251	0.819	0.582	0.588	**0.043**
Semi-Inf-Net & MC [27]	VGG16 + Res2Net	0.639	0.631	0.973	0.715	**0.904**	0.070	0.471	0.527	0.979	0.618	0.781	0.045
***FSS-2019-nCov**	Res2Net	**0.679**	**0.768**	0.980	**0.739**	0.894	**0.061**	**0.529**	**0.534**	**0.983**	**0.661**	**0.797**	0.045

↑ denote 'higher is better', ↓ denote 'lower is better'

Table 4. The results of evaluating different comparative models on the CT-2 dataset.

Methods	DSC↑	Sens↑	Spec↑	S_α ↑	E_ϕ ↑	MAE↓
U-Net [9]	0.337	0.682	0.841	0.523	0.649	0.221
H-DenseUNet [11]	0.419	0.635	0.964	0.547	0.561	0.167
U-Net++ [12]	0.462	0.881	0.937	0.589	0.614	0.115
SegNet [13]	0.453	0.844	0.932	0.624	0.6330	0.107
Inf-Net [27]	0.579	0.870	0.974	0.651	0.742	0.054
SE-Net [43]	0.555	0.837	0.924	0.673	0.713	0.054
Semi-Inf-Net [27]	0.597	0.865	0.977	0.723	0.792	0.037
***FSS-2019-nCov**	**0.632**	**0.892**	**0.975**	**0.764**	**0.824**	**0.031**

↑ denote 'higher is better', ↓ denote 'lower is better'

improvements on DSC, Sensitivity and S_α respectively over the best result in each measure. On the other hand, for consolidation lesion, the FSS-2019-nCov achieved **0.529** of DSC, **0.534** of Sens, **0.983** of Spec, **0.661** of S_α, **0.797** of E_ϕ, and 0.045 of MAE. It is observed that the model has similar behavior in segmenting this lesion, as noted from results in Table 3 where we attain 5%, 1%,1%, and 5% improvement on DSC, Sensitivity, Specificity, and S_α, respectively. However, Semi-Inf-Net-FCN8s obtained a slight improvement over our model for the MEA measure, which could result from the effectiveness of parallel partial decoders in pixelwise error between the segmentation result and GT even if they increase computation burden. The above discussion further validates that integrating TL, SSL, FSL in a single segmentation framework extensively improves the segmentation performance is scarce annotation scenarios.

E. Generalization analysis

The generalization capability of any segmentation model is an important aspect to demonstrate its effectiveness in real-world scenarios. In view of this, to understand and analyze the generalization capability of the proposed FSS-2019-nCov, we propose to evaluate it against previously mentioned comparative studies on the CT-2 data and present the corresponding results presented in Table 4. It can be noted that the proposed FSS-2019-nCov has a robust generalization performance overcoming all other approaches on all measures even though the data comprises axial slices with no lesions (i.e., lesion-free slice). This might be reasoned by utilizing two datasets during training, i.e., CT-1 data and unannotated CT slice extracted from 20 CT volumes. Further, the unannotated data comprises many lesion-free slices with no lesion to assure that FSS-2019-nCov can efficiently handle deal with lesion-free slices. Therefore, we can conclude that FSS-2019-nCov is a general lesion segmentation technique that can be applied to a variety of diseases.

F. Ablation Experiment

1) *Impact of RR module*

In this part, we inspect the ideal positions of RR blocks for smoothing knowledge interactions between the conditioner path and the segmentation path and also compare the FSS-2019-nCov performance when using recombination block only, recalibration block, and both blocks together (RR). Meanwhile, this experiment seeks to find the position and the type of interaction blocks—here, we fix all the network parameters and they later analyzed in subsequent sections. With two types of interaction blocks (i.e., SegSE, and recombination) and four possible positions for interaction block, there are twelve model variants termed as BLK-1, BLK-2.etc. In Table 5, we provide the segmentation DSC performance in terms of whole lung scenario and multi-class scenario for every configuration in these twelve model variants. It could be noted that BLK-3, 6, 9, 12 with Recombination and Recalibration (RR) blocks (the ones that have ✓ under the R (SegSE), and R column) yield the highest DSC score, which demonstrates the efficiency of RR interaction modules in effectuating the interactions between two paths of FSS-2019-nCov architecture. This network behavior could be explained due to concurrent spatial and channel squeezing using $Conv\ 1 \times 1$ to reduce the number of feature maps and increase their number later hence empower their representational power to convey the relevant information from the conditioner path to the segmentation path. Additionally, we could observe that the BLK-12 with RR blocks between all encoder, CE, and decoder blocks, BLK-12 attained the maximum DSC since it achieved a 3% improvement for infection segmentation over the best DSC obtained by other variants that correspond to BLK-11. In the multi-class scenario, BLK-12 attained 1% and 2% improvements over GGO and consolidation correspondingly. This improvement is potentially associated with the complexity and size of each class. In other words, the size and contrast of the GGO facilitate its segmentation in comparison to consolidation. Also, BLk-1: BLK-9 show poor performance in comparison to BLK-10: BLK-12. This shows that extra interactions enable better learning. It is obviously notable that

Table 5. Comparison between a different variant of the model to investigate the optimal position and kind of interaction blocks

	Position of RR Block			Interaction block		DSC		
	Enc	CE	Dec	R (SegSE)	R	Infection	GGO	Cons
BLK-1	✓	×	×	✓	×	0.661	0.475	0.405
BLK-2	✓	×	×	×	✓	0.414	0.274	0.314
BLK-3	✓	×	×	✓	✓	0.698	0.513	0.426
BLK-4	×	✓	×	✓	×	0.571	0.369	0.321
BLK-5	×	✓	×	×	✓	0.327	0.221	0.221
BLK-6	×	✓	×	✓	✓	0.545	0.373	0.395
BLK-7	×	×	✓	✓	×	0.623	0.441	0.234
BLK-8	×	×	✓	×	✓	0.421	0.239	0.326
BLK-9	×	×	✓	✓	✓	0.644	0.455	0.361
BLK-10	✓	✓	✓	✓	×	0.733	0.669	0.511
BLK-11	✓	✓	✓	×	✓	0.77	0.632	0.514
BLK12	✓	✓	✓	✓	✓	**0.798**	**0.679**	**0.529**

R (SegSE) represent recalibration block,R represent recombination block

model variants with encoder interactions (i.e., BLK-1, 2, 3) show higher performance compared to model variants with decoder interactions (i.e., BLK-7: BLK-9). This shows that encoder interactions are much representative and influential than CE or decoder interactions. Nevertheless, as BLK-12 yielded better performance than other model configurations. This could be explained by the encoder and decoder interactions generating complementary knowledge representation to the segmentation path to enable more enhanced segmentation of the query slices. From these discussions, it could be deduced that applying RR blocks at Encoder, CE, decoder leads to better performance than applying them to any single position.

2) *Impact of Skip Connection*

The connections have been regarded as a principle design choice in most F-CNN. It enables the concatenation of encoder output map and input feature maps of the decoder block with the same spatial resolution. This connection helps the decoder in capturing the contextual information and hence smooths the flow of gradient. In light of this, we start building our model by applying skip connections in both the conditioner path and segmentation path, and the result show copy over effect [43]. This means that the prediction on the query slice is almost symmetric to the support mask despite the difference between the support and query slice. Therefore, we conducted several experiments to investigate the impact of using skip connection on model performance in terms of DSC and hence on the copy over effect. In this experiment, we fixed all network parameters used in BLK-12 and just try different skip connection configurations. Thus, the performance of FSS-2019-nCov with and without skip connections is presented in Table 6. It could be noted that the DSC of whole infection segmentation decreased by 4% and also decreased by 3% in the case of GGO and Consolidation when applying skip connections in the two paths of the network (i.e., conditioner and segmentation paths). Also, applying skip connection on only the segmentation path obviously yields unsatisfactory results. Moreover, including the skip connections in the conditioner path results in a 5% decrease in DSC in different segmentation scenarios.

Table 6. Experimental results for analyzing the impact of using sip connection in E-Darchitecture.

Skip Connections		DSC		
Conditioner path	Segmentation path	infection	GGO	Cons
✓	×	0.749	0.573	0.485
✓	✓	0.752	0.644	0.506
×	×	**0.798**	**0.679**	**0.529**
×	✓	0.415	0.256	0.201

Table 7. Ablation experiments on the proposed FSS-2019-nCov on CT-1 dataset.

Methods	DSC↑	Sens↑	Spec↑	S_α ↑	E_ϕ ↑	MAE↓
Baseline w/o pretraining	0.643	0.681	0.834	0.721	0.719	0.278
Baseline w/ pre-training	0.665	0.718	0.881	0.741	0.735	0.181
Backbone + SAC (atrous)	0.701	0.737	0.929	0.769	0.815	0.166
Backbone + SAC (SS-Conv)	0.731	0.748	0.956	0.781	0.863	0.105
Backbone + MPP	0.715	0.712	0.941	0.749	0.841	0.119
***FSS-2019-nCov**	**0.798**	**0.803**	**0.986**	**0.834**	**0.908**	**0.065**

3) *Impact of pre-training*

In this experiment, we choose U-Net with a non-pre-trained encoder as a baseline architecture for both segmentation and conditioner paths. Then we replace the baseline encoder with a pre-trained one to obtain enhanced performance. The architecture with a pre-trained residual encoder is called the 'Backbone'. The result with and without pre-training compared and it could be noted that using pre-trained Res2Net clearly improves performance as depicted in Table 7.

4) *Impact of SAC module*

The proposed SAC block utilizes a variety of SS-Conv organized in the form of an Inception module to extract high-level spatial representation. Thus, to investigate the effectiveness of SS- Conv, we used atrous convolution to replace the SS-Conv in the SAC block (denoted Backbone + SAC (atrous)). Table 7 shows that the proposed SAC block achieves 3% DSC improvement over the traditional atrous block (Backbone + SAC (atrous)) and reduces the MAE with 0.061 in whole infection segmentation to achieve a similar improvement in other metrics. This, in turn, demonstrates that SS-Conv effectively enables improved feature fusion to extract high-level multi-scale contextual feature maps with high resolution and hence improve segmentation performance.

5) *Impact of MPP*

In an attempt to validate the usefulness of the proposed MPP block, we experiment with our Backbone architecture with and without MPP blocks for infection segmentation, as presented in Table 7. It is obviously noted that the MPP block boosts the model performance. The 'Backbone + MPP' achieved a 5% improvement on DSC, and reduced the MEA with 0.057. This indicates MPP block could effectively encode the local contextual representation from the encoder generated maps feature maps.

Table 8. The results of evaluating the proposed FSS-2019-nCovon CT-1 using different learning paradigms.

Methods	Semi-supervised learning	Supervised learning
DSC↑	0.798	0.679
Sens↑	0.803	0.744
Spec↑	0.986	0.959
S_α ↑	0.834	0.774
E_ϕ ↑	0.908	0.803
MAE↓	0.065	0.105

↑ denote 'higher is better', ↓ denote 'lower is better'

6) *Impact of Semi-supervised training*

In order to demonstrate the efficiency of semi-supervised training of the proposed FSS-2019-nCov, we compare performance when trained in a supervised and semi-supervised manner, and we report the corresponding results in Table 8. It can be noted that semi-supervised training shows significant performance improvements in segmenting infection lesion (i.e. DSC of 0.119, Sensitivity of 0.059, Specificity of 0.027, S_α **of** 0.06, E_ϕ **of** 0.105, and **MAE of** 0.040. This observation provides clear evidence regarding the effectiveness of incorporating unannotated CT data for training FSS-2019-nCov.

V. MANAGERIAL IMPLICATIONS

COVID-19 segmentation is the task of determining the infection area within lung CT scans. This task could be addressed as a binary classification problem or a multi-classification problem. In binary classification scenarios, we aim to distinguish between infected and uninfected areas. In a multi-class scenario, we aim to distinguish between different types of infection. The key challenge of this study is the limited amount of labeled CT scans. We propose a novel architecture that integrates pre-trained encoder, FSS, and SSL to overcome this limitation. The Res2Net50-based encoder enables improved network convergence. The FSS architecture enables learning from limited support samples and better generalization of query samples. We introduce adaptive recombination and recalibration module between the correspondence positions in the conditioner and segmentation path to facilitate knowledge representation exchange. This is established by our experiments since it can be safely claimed that RR significantly finetune knowledge interaction and hence improve the performance. Meanwhile, the CE module enables capturing contextual information of infection at different scales, facilitating the detection of different sizes of infections. Comprehensive experiments confirmed the effectiveness of each block. As a direct implication, the proposed FSS-2019-nCov in study work can be utilized to develop an automated lung infection segmentation system with scarcely annotated data.

VI. SHORTCOMINGS AND POSSIBLE REMEDIES

Extra deep learning improvement will be addressed by future work in terms of performance improvement and computational complexity reduction. We aim to investigate three crucial challenges that we regard as specifically related to the medical image analysis community. (1) The training configuration of FSS-2019-nCov denotes a challenging task since it still necessitates a comprehensive parameter improving to attain the highest results. An automatized tuning tool can be used for this. (2) The predictions usually lack laborious uncertainty quantification. We aim to develop Bayesian variants or fuzzified variants of proposed FSS-2019-nCov that could enable estimating uncertainty in prediction. (3) Although extensive analysis has provided us with a great understanding of the behavior of FSL and FSS, accountability and interpretability are considered as a downside of our FSS-2019-nCov and an attention technique could mitigate this.

VII. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel semi-supervised few-shot segmentation model for COVID-19 segmentation from axial CT scans using dual-path architecture. The two paths had a symmetric structure and comprise an encoder-decoder architecture with a smoothed context fusion module. The encoder architecture was based on pre-trained ResNet34 architecture to facilitate the learning process. We proposed to merge recombination and recalibration to transfer learned knowledge from the support set to be used for query slices segmentation. The model trained in semi-supervised strategy by incorporating unlabeled CT slices and labeling one during training, improving generalization performance. We investigated the proposed FSS-2019-nCov and numerous baselines on publicly available COVID-19 CT scans. The results showed that our model could outperform all approaches to multiple evaluation metrics. We also introduced comprehensive experiments for architectural selection concerning RR blocks, Skip connections, and the proposed building blocks. However, the segmentation performance of the proposed FSS-2019-nCov was unable to achieve a very precise segmentation due to limited supervision, which could be handled with a generative learning schema. An additional limitation was a lack of volumetric data representation, which could be alleviated by expanding our model to 3D CT volumes of COVID-19. Consequently, we aim to investigate the segmentation of COVID-19 using a large amount of volumetric 3D data in the near future.

Conflict of interest

The authors declare that there is no conflict of interest in the research.

Funding

This work is partly supported by VC Research (VCR 0000088).

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

REFERENCES

[1] C. A. Devaux, J.-M. Rolain, P. Colson, and D. Raoult, "New insights on the antiviral effects of chloroquine against coronavirus: what to expect for COVID-19?," International Journal of Antimicrobial Agents, p. 105938, 2020.

[2] F. Song, N. Shi, F. Shan, Z. Zhang, J. Shen, H. Lu, et al., "Emerging 2019 Novel Coronavirus (2019-nCoV) Pneumonia," Radiology, vol. 295, pp. 210-217, 2020.

[3] F. Shan+, Y. Gao+, J. Wang, W. Shi, N. Shi, M. Han, et al., "Lung infection quantification of covid-19 in ct images with deep learning," arXiv preprint arXiv:2003.04655, 2020.

[4] Y. Fang, H. Zhang, J. Xie, M. Lin, L. Ying, P. Pang, *et al.*, "Sensitivity of chest CT for COVID-19: comparison to RT-PCR," *Radiology*, p. 200432, 2020.

[5] A. Bernheim, X. Mei, M. Huang, Y. Yang, Z. A. Fayad, N. Zhang, et al., "Chest CT findings in coronavirus disease-19 (COVID-19): relationship to duration of infection," Radiology, p. 200463, 2020.

[6] D. Caruso, M. Zerunian, M. Polici, F. Pucciarelli, T. Polidori, C. Rucci, et al., "Chest CT features of COVID-19 in Rome, Italy," Radiology, p. 201237, 2020.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431-3440.

[8] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[9] 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*, 2015, pp. 234-241.

[10] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, *et al.*, "Attention gated networks: Learning to leverage salient regions in medical images," *Medical image analysis*, vol. 53, pp. 197-207, 2019.

[11] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes," *IEEE transactions on medical imaging*, vol. 37, pp. 2663-2674, 2018.

[12] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning Skip Connections to Exploit Multi-scale Features in Image Segmentation," *IEEE Transactions on Medical Imaging*, 2019.

[13] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, pp. 2481-2495, 2017.

[14] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801-818.

[15] J. Guan, Z. Lu, T. Xiang, A. Li, A. Zhao, and J.-R. Wen, "Zero and few shot learning with semantic feature synthesis and competitive learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[16] A. Shaban, S. Bansal, Z. Liu, I. Essa, and B. Boots, "One-shot learning for semantic segmentation," *arXiv preprint arXiv:1709.03410*, 2017.

[17] K. Rakelly, E. Shelhamer, T. Darrell, A. A. Efros, and S. Levine, "Few-shot segmentation propagation with guided networks," *arXiv preprint arXiv:1806.07373*, 2018.

[18] Z. Cao, T. Zhang, W. Diao, Y. Zhang, X. Lyu, K. Fu, *et al.*, "Meta-Seg: A Generalized Meta-Learning Framework for Multi-Class Few-Shot Semantic Segmentation," *IEEE Access*, vol. 7, pp. 166109-166121, 2019.

[19] Y. Tong, J. K. Udupa, D. Odhner, C. Wu, S. J. Schuster, and D. A. Torigian, "Disease quantification on PET/CT images without explicit object delineation," *Medical image analysis*, vol. 51, pp. 169-183, 2019.

[20] A. Kumar, M. Fulham, D. Feng, and J. Kim, "Co-learning feature fusion maps from PET-CT images of lung cancer," *IEEE Transactions on Medical Imaging*, vol. 39, pp. 204-217, 2019.

[21] O. Ozdemir, R. L. Russell and A. A. Berlin, "A 3D Probabilistic Deep Learning System for Detection and Diagnosis of Lung Cancer Using Low-Dose CT Scans," in *IEEE Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1419-1429, May 2020, doi: 10.1109/TMI.2019.2947595.

[22] S. Gerard, T. Patton, G. Christensen, J. Bayouth, and J. Reinhardt, "FissureNet: A Deep Learning Approach For Pulmonary Fissure Detection in CT Images," *IEEE transactions on medical imaging*, vol. 38, pp. 156-166, 2019.

[23] J. Jiang, Y. Hu, C. Liu, D. Halpenny, M. Hellmann, J. Deasy, *et al.*, "Multiple Resolution Residually Connected Feature Streams for Automatic Lung Tumor Segmentation From CT Images," *IEEE transactions on medical imaging*, vol. 38, pp. 134-144, 2019.

[24] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical image analysis*, vol. 54, pp. 280-296, 2019.

[25] N. Kumar *et al.*, "Hyperspectral Tissue Image Segmentation Using Semi-Supervised NMF and Hierarchical Clustering," in *IEEE Transactions on Medical Imaging*, vol. 38, no. 5, pp. 1304-1313, May 2019, doi: 10.1109/TMI.2018.2883301.

[26] S. Y. Shin, S. Lee, I. D. Yun, S. M. Kim and K. M. Lee, "Joint Weakly and Semi-Supervised Deep Learning for Localization and Classification of Masses in Breast Ultrasound Images," in *IEEE Transactions on Medical Imaging*, vol. 38, no. 3, pp. 762-774, March 2019, doi: 10.1109/TMI.2018.2872031.

[27] D.-P. Fan, T. Zhou, G.-P. Ji, Y. Zhou, G. Chen, H. Fu, *et al.*, "Inf-Net: Automatic COVID-19 Lung Infection Segmentation from CT Scans," *arXiv preprint arXiv:2004.14133*, 2020.

[28] J. Liang, R. He, Z. Sun, and T. Tan, "Exploring uncertainty in pseudo-label guided unsupervised domain adaptation," *Pattern Recognition*, vol. 96, p. 106996, 2019.

[29] F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, *et al.*, "Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19," *IEEE Reviews in Biomedical Engineering*, 2020.

[30] V. Rajinikanth, N. Dey, A. N. J. Raj, A. E. Hassanien, K. Santosh, and N. Raja, "Harmony-search and otsu based system for coronavirus disease (COVID-19) detection using lung CT scan images," *arXiv preprint arXiv:2004.03431*, 2020.

[31] S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, *et al.*, "A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19)," *MedRxiv*, 2020.

[32] J. Chen, L. Wu, J. Zhang, L. Zhang, D. Gong, Y. Zhao, *et al.*, "Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study," *medRxiv*, 2020.

[33] A. Narin, C. Kaya, and Z. Pamuk, "Automatic Detection of Coronavirus Disease (COVID-19) Using X-ray Images and Deep Convolutional Neural Networks," *arXiv preprint arXiv:2003.10849*, 2020.

[34] Y. Song, S. Zheng, L. Li, X. Zhang, X. Zhang, Z. Huang, et al., "Deep learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) with CT images," medRxiv, 2020.

[35] W. Shi, X. Peng, T. Liu, Z. Cheng, H. Lu, S. Yang, *et al.*, "Deep Learning-Based Quantitative Computed Tomography model in Predicting the Severity of COVID-19: A Retrospective Study in 196 Patients," 2020.

[36] F. Shi, L. Xia, F. Shan, D. Wu, Y. Wei, H. Yuan, *et al.*, "Large-scale screening of covid-19 from community acquired pneumonia using infection size-aware classification," *arXiv preprint arXiv:2003.09860*, 2020.

[37] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H.Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[38] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, *et al.*, "CE-Net: context encoder network for 2D medical image segmentation," *IEEE transactions on medical imaging*, vol. 38, pp. 2281-2292, 2019.

[39] Z. Wang and S. Ji, "Smoothed Dilated Convolutions for Improved Dense Prediction," *arXiv preprint arXiv:1808.08931*, 2018.

[40] A.-M. Rickmann, A. G. Roy, I. Sarasua, and C. Wachinger, "Recalibrating 3D ConvNets with Project & Excite," *IEEE Transactions on Medical Imaging*, 2020.

[41] S. Pereira, A. Pinto, J. Amorim, A. Ribeiro, V. Alves, and C. A. Silva, "Adaptive feature recombination and recalibration for semantic segmentation with Fully Convolutional Networks," *IEEE transactions on medical imaging*, vol. 38, pp. 2914-2925, 2019.

[42] "COVID-19 CT segmentation dataset," https://medicalsegmentation.com/covid19/, accessed: 2020-04-11.

[43] A. G. Roy, S. Siddiqui, S. Pölsterl, N. Navab, and C. Wachinger, "'Squeeze & excite'guided few-shot segmentation of volumetric images," *Medical image analysis*, vol. 59, p. 101587, 2020.

[44] 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)*, 2016, pp. 565-571.

[45] C. Baur, S. Albarqouni, and N. Navab, "Auxiliary Manifold Embedding for Fully Convolutional Networks," *arXiv preprint arXiv:1703.06000*, 2017.

[46] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, pp. 834-848, 2017.

[47] S. Yang, G. Lin, Q. Jiang, and W. Lin, "A dilated inception network for visual saliency prediction," *IEEE Transactions on Multimedia*, 2019.

[48] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting More Attention to Video Salient Object Detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8546-8556.

[49] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 221-230.
[50] N. Dong and E. Xing, "Few-Shot Semantic Segmentation with Prototype Learning," in *BMVC*, 2018.
[51] K. Rakelly, E. Shelhamer, T. Darrell, A. A. Efros, and S. Levine, "Few-shot segmentation propagation with guided networks," *arXiv preprint arXiv:1806.07373*, 2018.
[52] A. K. Mondal, J. Dolz, and C. Desrosiers, "Few-shot 3d multi-modal medical image segmentation using generative adversarial learning," *arXiv preprint arXiv:1810.12241*, 2018.
[53] [1] A. Zhao, G. Balakrishnan, F. Durand, J. V. Guttag, and A. V. Dalca, "Data augmentation using learned transforms for one-shot medical image segmentation," *arXiv preprint arXiv:1902.09383*, 2019.
[54] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning" In AAAI, vol. 4, 2017, p. 12.
[55] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881-2890.
[56] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in ICCV, 2017, pp. 4548–4557.
[57] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking RGB-D Salient Object Detection: Models, Data Sets, and Large-Scale Benchmarks," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.
[58] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-Net: Going deeper with nested U-structure for salient object detection," *Pattern Recognition*, vol. 106, p. 107404, 2020.
[59] J. Ma, Y. Wang, X. An, C. Ge, Z. Yu, J. Chen, *et al.*, "Towards Efficient COVID-19 CT Annotation: A Benchmark for Lung and Infection Segmentation," *arXiv preprint arXiv:2004.12537*, 2020.
[60] S. Mittal, M. Tatarchenko, Ö. Çiçek, and T. Brox, "Parting with Illusions about Deep Active Learning," *arXiv preprint arXiv:1912.05361*, 2019.
[61] X. Wang et al., "A Weakly-Supervised Framework for COVID-19 Classification and Lesion Localization From Chest CT," in IEEE Transactions on Medical Imaging, vol. 39, no. 8, pp. 2615-2625, Aug. 2020, doi: 10.1109/TMI.2020.2995965.
[62] Z. Zhou, Z. He, and Y. Jia, "AFPNet: A 3D Fully Convolutional Neural Network with Atrous-convolution Feature Pyramid for Brain Tumor Segmentation via MRI Images," *Neurocomputing*, 2020.
[63] X. Lian, Y. Pang, J. Han, and J. Pan, "Cascaded hierarchical atrous spatial pyramid pooling module for semantic segmentation," *Pattern Recognition*, vol. 110, p. 107622, 2020.