Title

CovSegNet: A Multi Encoder-Decoder Architecture for Improved Lesion Segmentation of COVID-19 Chest CT Scans.

Journal

IEEE Transactions on Artificial Intelligence, 2(3)

Authors

Mahmud, Tanvir

Rahman, Md

Fattah, Shaikh

et al.

Publication Date

2021-06-01

DOI

10.1109/TAI.2021.3064913

Peer reviewed

CovSegNet: A Multi Encoder–Decoder Architecture for Improved Lesion Segmentation of COVID-19 Chest CT Scans

Tanvir Mahmud, *Student Member, IEEE*, Md Awsafur Rahman, *Student Member, IEEE*, Shaikh Anowarul Fattah, *Senior Member, IEEE*, and Sun-Yuan Kung, *Life Fellow, IEEE*

***Abstract*—Automatic lung lesion segmentation of chest computer tomography (CT) scans is considered a pivotal stage toward accurate diagnosis and severity measurement of COVID-19. Traditional U-shaped encoder–decoder architecture and its variants suffer from diminutions of contextual information in pooling/upsampling operations with increased semantic gaps among encoded and decoded feature maps as well as instigate vanishing gradient problems for its sequential gradient propagation that result in suboptimal performance. Moreover, operating with 3-D CT volume poses further limitations due to the exponential increase of computational complexity making the optimization difficult. In this article, an automated COVID-19 lesion segmentation scheme is proposed utilizing a highly efficient neural network architecture, namely CovSegNet, to overcome these limitations. Additionally, a two-phase training scheme is introduced where a deeper 2-D network is employed for generating region-of-interest (ROI)-enhanced CT volume followed by a shallower 3-D network for further enhancement with more contextual information without increasing computational burden. Along with the traditional vertical expansion of Unet, we have introduced horizontal expansion with multistage encoder–decoder modules for achieving optimum performance. Additionally, multiscale feature maps are integrated into the scale transition process to overcome the loss of contextual information. Moreover, a multiscale fusion module is introduced with a pyramid fusion scheme to reduce the semantic gaps between subsequent encoder/decoder modules while facilitating the parallel optimization for efficient gradient propagation. Outstanding performances have been achieved in three publicly available datasets that largely outperform other state-of-the-art approaches. The proposed scheme can be easily extended for achieving optimum segmentation performances in a wide variety of applications.**

***Impact Statement*—With lower sensitivity (60–70%), elongated testing time, and a dire shortage of testing kits, traditional RTPCR based COVID-19 diagnostic scheme heavily relies on postCT based manual inspection for further investigation. Hence, automating the process of infected lesions extraction from chestCT volumes will be major progress for faster accurate diagnosis of COVID-19. However, in challenging conditions with diffused, blurred, and varying shaped edges of COVID-19 lesions, conventional approaches fail to provide precise segmentation of lesions that can be deleterious for false estimation and loss of information. The proposed scheme incorporating an efficient neural network architecture (CovSegNet) overcomes the limitations of traditional approaches that provide significant improvement of performance (8.4% in averaged dice measurement scale) over two datasets. Therefore, this scheme can be an effective, economical tool for the physicians for faster infection analysis to greatly reduce the spread and massive death toll of this deadly virus through mass-screening.**

***Index Terms*—Artificial intelligence (AI), biomedical imaging, computer aided analysis, image segmentation, neural networks.**

I. Introduction

WITH the recent outbreak of Coronavirus disease-2019 (COVID-19), the world has experienced an unprecedented number of deaths with a major collapse in the healthcare system throughout the world [1], [2]. Early diagnosis is the primary concern to control this global pandemic at this stage for its extreme infectious nature [3]. Though reverse transcription-polymerase chain reaction is considered as the gold standard for diagnosing COVID-19, its longer time requirement, lower sensitivity with a massive shortage of test-kits have already engendered the extreme urgency of alternative automated diagnostic schemes [4], [5]. Due to the wide applicability of the artificial intelligence (AI) tools in numerous clinical diagnostic measures, it has enormous potential to expedite the diagnostic process of COVID-19 through automated analysis and interpretation of the clinical record [6], [7].

Chest radiography has already been proven to be an effective source for COVID diagnostics due to its major implications relating to various levels of lung infections [8]. Computer tomography (CT) scan and chest X-ray have been extensively explored in the literature to establish an automated AI-based COVID diagnostic scheme [9]–[11]. Despite the easier access to chest X-ray, CT scans are more widely accepted due to its finer details leveraging the accurate diagnosis of COVID infections. Precise segmentation of lung lesions in chest CT scans is one of the most demanding and challenging aspects for faster diagnosis of COVID-19 due to the shortage of annotated data, diverse levels of infections, and novel types and characteristics of the infections [12].

Processing 3-D CT volume at a whole increases computational complexity exponentially that makes the optimization and

Manuscript received November 30, 2020; revised January 8, 2021 and January 23, 2021; accepted March 1, 2021. Date of publication March 15, 2021; date of current version September 14, 2021. This article was recommended for publication by Associate Editor Supratik Mukhopadhyay upon evaluation of the reviewers' comments. *(Corresponding author: Shaikh Anowarul Fattah.)*

Tanvir Mahmud, Md Awsafur Rahman, and Shaikh Anowarul Fattah are with the Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Dhaka 1000, Bangladesh (e-mail: tanvirmahmud@eee.buet.ac.bd; mdawsafurrahman@ug.eee.buet.ac.bd; fattah@eee.buet.ac.bd).

Sun-Yuan Kung is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: kung@princeton.edu).

Digital Object Identifier 10.1109/TAI.2021.3064913

convergence more difficult limiting the architectural diversity of the network. The most widely used alternative of 3-D processing is to operate separately on 2-D slices extracted from the CT volume [12]–[16]. However, such slice-based processing loses interslice contextual information that results in suboptimal performance. In [17]–[20], smaller subvolumes are extracted from the original 3-D volumes to minimize the computational burden as well as to utilize 3-D contextual information. However, such methods suffer from intervolume contextual information loss by considering a smaller portion of the whole set at a time as well as increases complexity to process subvolume level prediction into the final result.

A wide variety of approaches have been introduced in recent years for segmenting the region-of-interest in diverse applications. In [21], a fully connected network is introduced that produces multiple scales of encoded feature maps and reconstructs the segmentation mask utilizing these encoded representations. In [22], Unet architecture is introduced by integrating an inverted decoder module following the encoder module to gradually reconstruct the mask that gains much popularity over the years. However, several architectural limitations of Unet are identified as follows that provides suboptimal performance.

1) The skip connection introduced in Unet generates semantic gap between corresponding feature scale of encoder–decoder modules, which mainly arises from the direct concatenation of two semantically dissimilar feature maps. As the encoder module encodes the input image gradually into more generalized feature representation, it contains richer details compared to the corresponding decoded feature map, which contains more information for the reconstruction of the final segmentation mask. These existing semantic gaps between corresponding encoder and decoder feature maps make the optimization process more difficult to converge for such direct concatenations through skip connections.
2) Contextual information loss occurs in traditional pooling/strided convolution-based downsampling operations that become more eminent with deeper architecture. Such downsampling operations are mainly carried out for generating more generalized, sparser feature representation with increased channels and reduced spatial resolution of the feature map. However, these operations also lead to loss of contextual information that greatly rises with the increase of vertical depth of the network. Similarly, the traditional upsampling operations fail to properly incorporate contextual information.
3) The vanishing gradient problem rises in a deeper structure for sequential optimization of multiscale features. This problem mainly arises from the difficulty of gradient propagation through the deep stack of convolutional layers. Along with the incorporation of additional levels in the encoder and decoder stacks to make the network deeper, it becomes increasingly difficult to backpropagate the gradients through these levels for propagating through longer sequential paths that make the optimization of the deeper layers more difficult. Hence, this problem reduces the effective contributions of the deeper layers of the encoder and decoder modules for improper optimization.
4) Simplistic sequential convolutional layers are integrated into each level of encoder/decoder modules that lack enough architectural diversity to extract features from a broader spectrum, which is mainly caused by the linear propagation of gradients that reduces the impact of prior convolutional layers at each level for diminishing gradients. It lacks opportunity for the proper reuse of extracted features in the successive convolutions and lacks parallelism among convolutional layers required for better optimization, which lower the diversity of features generated at different levels of the network.

Different architectural modifications have been explored in recent years to overcome some of these limitations. To increase the diversity of operations at each scale of feature maps, numerous established network building blocks are integrated in encoder/decoder module, e.g., residual block [23], dense block [24], inception block [25], dilated residual block [26], and multires block [27]. To reduce the semantic gap between a particular scale of encoder and decoder, a residual path is proposed in MultiResUnet architecture instead of a direct skip connection of Unet [27]. However, the semantic gap generated between multiscale feature maps of encoder and decoder modules still persists. In Unet++ [28], a nested stack of convolutional layers is introduced to reduce the semantic gaps. But, it increases computational complexity considerably, which makes convergence difficult. In [19], Vnet is proposed that utilizes residual building blocks in Unet architecture, whereas in [20], cascaded-Vnet is presented for performance improvement that utilizes a dual-stack of the cascaded encoder–decoder module. Nevertheless, with existing numerous architectural limitations of traditional U-shaped architecture in each stage, it increases semantic gaps with the additional encoding–decoding stage as well as increases vanishing gradient issues with contextual information loss that open up opportunities for further optimization.

In this article, an improved, automated scheme is proposed for precise lesion segmentation of COVID-19 chest CT volumes by overcoming the limitations of traditional approaches with a novel deep neural network architecture, named as CovSegNet. The major contributions of this article are summarized as follows.

1) Along with the opportunity of vertical expansion, a horizontal expansion strategy is introduced in the CovSegNet architecture. In the vertical expansion mechanism, the encoder and decoder modules are deepened, whereas in horizontal expansion, several encoding–decoding stages are integrated. As discussed earlier, loss of contextual information occurs when the network is vertically expanded through subsequent downsampling operations, though it provides the opportunity for improved generalization through incorporating features from higher levels, whereas the horizontal expansion mechanism assists to integrate more detailed features at each level for finer reconstruction that helps to recover the loss of contextual information. As a result, it provides the opportunity

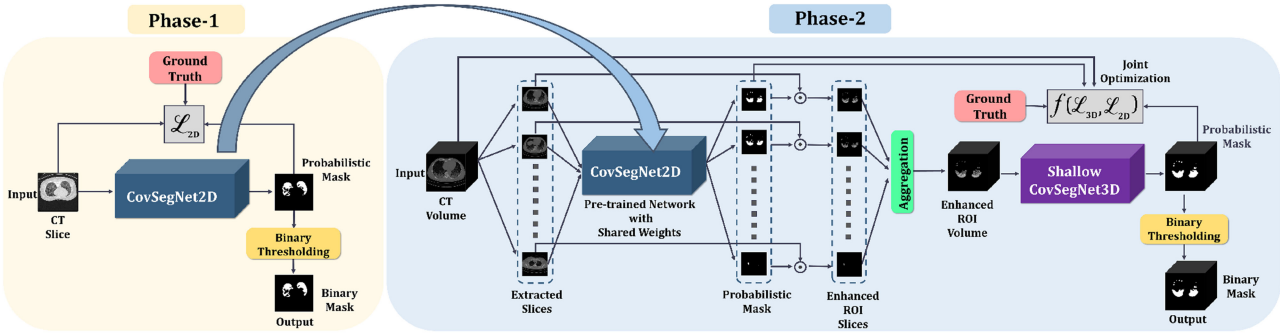

Fig. 1. Workflow of the proposed scheme for segmenting lung lesions of COVID-19 in CT volume. In phase-1, CovSegNet2D is trained and optimized with extracted 2D-CT-slices. In phase-2, this pretrained CovSegNet2D (obtained from phase-1) is fine-tuned for generating the ROI-enhanced CT volume, whereas a shallower form of CovSegNet3D is trained for more precise volumetric segmentation through the joint optimization.

to increase generalization while exploiting the available contextual information through an optimal combination of horizontal and vertical stages.

2) For further replenishing the loss of contextual information in traditional pooling/upsampling operations, a scale transition scheme is introduced in the encoder/decoder module by incorporating multiscale feature maps from preceding levels. This scale transition scheme also improves the gradient flow across different feature scales of a particular encoder/decoder module.
3) For reducing semantic gaps among corresponding feature scales of the encoder–decoder modules, a multiscale fusion (MSF) module is introduced in between successive encoder–decoder modules. This module fuses multiscale feature representations, generated at preceding encoder/decoder modules through pyramid fusion (PF) scheme, to generate representational features with reduced semantic gap and improved contextual information for the following decoder/encoder module, instead of directly connecting corresponding feature scales, such as Unet. Moreover, this module establishes parallel linkage among multiscale feature maps of subsequent encoder–decoder modules that greatly improve the gradient flow across the network and helps to reduce the vanishing gradient problem.
4) A multiphase training approach is introduced for integrating the advantages of both the 2-D and 3-D data processing scheme to reach the optimum performance. 2-D processing provides faster processing with lower memory consumption while losing interslice contextual information, whereas 3-D processing exploits both the intraslice and interslice contextual information while increasing the computational burden. The proposed multiphase training solves this problem by integrating a deeper variant of CovSegNet2D followed by a much shallower variant of CovSegNet3D for exploiting all possible information while limiting the computational burden.
5) The proposed CovSegNet architecture is designed in a modular and structured way that can be adapted to its lightweight, shallow form to reduce complicacy with considerable performance as well as can be made very deep to increase diversity for incorporating finer details. This generic design provides more flexibility for tuning the design parameters in a wide variety of applications.
6) Extensive experimentations have been carried out to validate the effectiveness of the proposed scheme on two publicly available datasets containing chest CT scans from COVID-19 patients. Moreover, to validate the wide applicability of the proposed architecture, experimental results on a challenging, nonclinical, semantic segmentation dataset are also provided.

II. Methodology

The proposed scheme splits the segmentation of CT volumes into two subsequent phases to reduce the computational complexity of 3-D convolution as well as to take the advantages of multiscale 2-D convolutions (see Fig. 1). In the first phase of training, 2-D slices are extracted from the 3-D CT volumes and these are used for the optimization of CovSegNet2D (i.e., 2-D variant of the proposed CovSegNet architecture) from randomized initial state. After the optimization, the trained CovSegNet2D is capable of extracting lesions from 2-D slices. However, slice-based processing of input CT volumes will lead to loss of interslice contextual information resulting in suboptimal performance. Nevertheless, 2-D processing are computationally efficient and easy to optimize compared to the complete 3-D processing. To introduce further optimization for integrating the interslice contextual information of particular CT volume, phase-2 of the training stage is incorporated. Here, a hybrid volumetric processing scheme is introduced where the CovSegNet2D is initialized with the pretrained weights obtained from the phase-1 of the training. Thus, the complete 3-D CT volume is split into several 2-D slices that are processed through the CovSegNet2D to extract the region-of-interest in the 2-D CT-slices. Afterward, these enhanced 2-D CT-slices are aggregated to generate the region-of-interest (ROI)-enhanced CT volume where most of the redundant parts are suppressed. Nevertheless, to extract the interslice contextual information for further optimization, a lighter variant of CovSegNet3D is

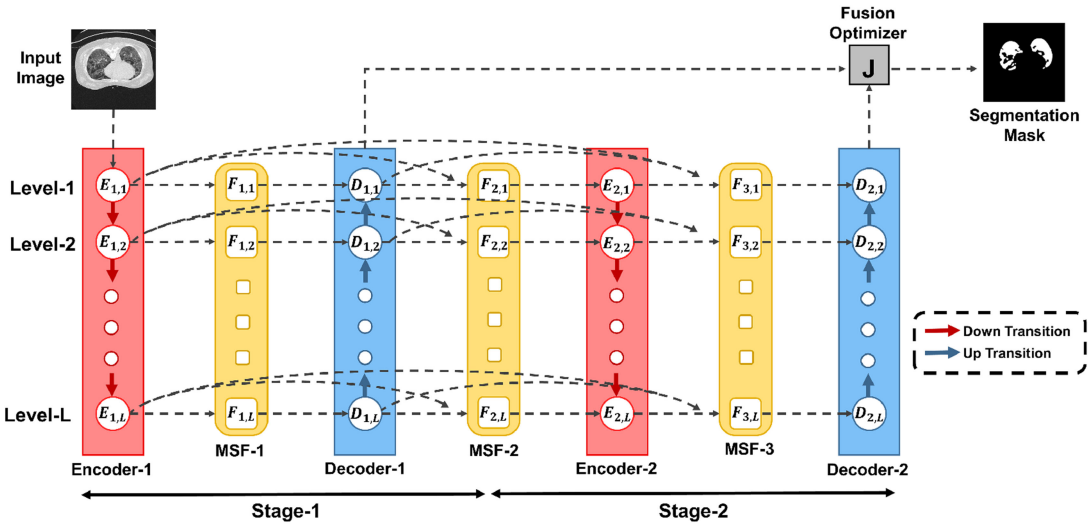

Fig. 2. Schematic representation of the two-stage implementation of the proposed CovSegNet architecture where two sequential encoder–decoder operational stages are employed with L subsequent levels. Three MSF modules are integrated in between subsequent modules. Generated feature maps from two decoder modules are optimized using the fusion optimizer. These encoder, decoder, and MSF modules are composed of several operational unit cells. Each unit of the MSF module integrates all scales of feature maps to generate the output fusion vector (see Fig. 6).

incorporated to operate on the ROI-enhanced CT volume. In the second phase of the training, CovSegNet3D will be optimized from scratch to extract the interslice contextual information, whereas CovSegNet2D will be fine-tuned for better extraction of the intraslice features. Hence, this joint optimization operation in phase-2 is supposed to optimize a very lighter variant of CovSegNet3D (as it operates on the ROI-enhanced volume), which reduces the computational burden of complete 3-D processing with very deep network. Moreover, as the CovSegNet2D is initially pretrained in the phase-1 for efficient 2-D slicewise processing, it greatly reduces the optimization complexity in phase-2 through generating ROI-enhanced CT volume. Hence, this hybrid networking scheme is capable of utilizing both the interslice and intraslice contextual information while greatly reducing the computational complexity of complete 3-D processing.

A. Problem Formulation

Let consider the set of CT volumes as $\mathbf{X}$, and their corresponding ground truths as $\mathbf{Y}$, such that $X_i \in \mathbb{R}^{h\times w\times s\times c}$, $Y_i \in \mathbb{R}^{h\times w\times s\times c}$, and $i = \{1, 2, 3, \ldots, N\}$, where (h, w, s, c) denote height, width, number of slices, and channels per slice, respectively, of a particular CT volume from total N number of CT volumes. Moreover, let consider $\mathbf{x_{i,j}} \in R^{h\times w\times c}$ as the ith slice from total S slices of jth CT volume and $\mathbf{y_{i,j}} \in R^{h\times w\times c}$ as its corresponding mask, such that $i = \{1, 2, \ldots, S\}$, and $j = \{1, 2, \ldots, N\}$. In the first phase of training, the objective function for slice-based optimization of CovSegNet2D is

$$\textbf{Phase1}: \text{argmin}_{\theta}\mathscr{L}_{2D}(\theta, \mathbf{y^P}, \mathbf{y}) \tag{1}$$

where θ denotes the network parameter of CovSegNet2D, $\mathbf{x}$, $\mathbf{y^P}$, and y denote the input 2-D slice, predicted probability mask, and corresponding ground truth mask, respectively.

In the phase-2 of training, the pretrained CovSegNet2D network obtained from phase-1 is employed to generate ROI-enhanced CT volume $\mathbf{X}'$, and thus

$$\mathbf{x}' = \mathbf{x} \odot \mathbf{y^P} \quad \forall\, \mathbf{x}' \in \mathbf{X},' \ \mathbf{x} \in \mathbf{X}, \ \mathbf{y^P} \in \mathbf{Y^P} \tag{2}$$

where $\odot$ denotes elementwise multiplication and $\mathbf{x}$ denotes 2-D CT slice, $\mathbf{x}'$ denotes ROI-enhanced CT-slice, and $\mathbf{y^P}$ denotes the predicted probability mask.

Afterward, optimization of the CovSegNet3D is carried out utilizing ROI-enhanced CT volume, whereas CovSegNet2D is fine-tuned to generate more accurate probability masks from 2-D slices, and the joint optimization objective function $\mathcal{F}$ can be formulated as

$$\textbf{Phase2}: \underset{\Theta_1,\Theta_2}{\text{argmin}}\, \mathcal{F}\{\mathscr{L}_{2D}(\Theta_1, \mathbf{y^P}, \mathbf{y}), \mathscr{L}_{3D}(\Theta_2, \mathbf{Y^P}, \mathbf{Y})\} \tag{3}$$

where Θ_1 denotes the network parameters of CovSegNet2D, Θ_2 denotes the network parameters of CovSegNet3D, and $\mathbf{X},' \mathbf{Y^P}$, and $\boldsymbol{Y}$ denote the ROI-enhanced CT volume, predicted 3-D mask, and corresponding 3-D ground truth, respectively.

B. Proposed CovSegNet Architecture

The proposed CovSegNet architecture is a generic representation of a network with a wide range of flexibility for increasing its applicability in different challenging conditions. This architecture can be designed for efficient operations in both 2-D and 3-D domains. Moreover, it can be made deeper/lighter according to the requirement of the applications.

In CovSegNet architecture, multiple stages of sequential encoding and decoding operations are carried out along with a fusion scheme of multiscale features in between subsequent encoder/decoder module. Each stage of the network consists of an encoder module and a corresponding decoder module. Hence,

TABLE I
ARCHITECTURAL AND OPERATIONAL DETAILS OF THE ENCODER, DECODER, AND MULTISCALE FUSION MODULES OF THE PROPOSED COVSEGNET2D FOR OPTIMUM PERFORMANCE IN INDEPENDENT SINGLE-NETWORK IMPLEMENTATION

Encoder module			Decoder module			Multi-scale fusion module		
Unit	**Ingredients**	**Output**	**Unit**	**Ingredients**	**Output**	**Unit**	**Ingredients**	**Output**
E-1	(Conv 1×1, Conv 3×3) × 4	512×512×16	D-5	(Conv 1×1, Conv 3×3) × 4	32×32×256	MSF-1	Upsample(2×2,4×4,8×8,16×16) Maxpool(2×2,4×4) Conv 1×1, Conv 3×3	512×512×16
DT-1	Conv 2×2, Stride 2	256×256×32	UT-4	Deconv 2×2, Stride 2	64×64×128			
E-2	(Conv 1×1, Conv 3×3) × 4	256×256×32	D-4	(Conv 1×1, Conv 3×3) × 4	64×64×128	MSF-2	Maxpool (2×2,4×4) Upsample(2×2,4×4,8×8) Conv 1×1, Conv 3×3	256×256×32
DT-2	Maxpool 2×2 Conv 2×2, Stride 2	128×128×64	UT-3	Upsample 2×2 Deconv 2×2, Stride 2	128×128×64			
E-3	(Conv 1×1, Conv 3×3) × 4	128×128×64	D-3	(Conv 1×1, Conv 3×3) × 4	128×128×64	MSF-3	Maxpool (2×2,4×4) Upsample(2×2,4×) Conv 1×1, Conv 3×3	128×128×64
DT-3	Maxpool (2×2, 4×4) Conv 2×2, Stride 2	64×64×128	UT-2	Upsample (2×2, 4×4) Deconv 2×2, Stride 2	256×256×32			
E-4	(Conv 1×1, Conv 3×3) × 4	64×64×128	D-2	(Conv 1×1, Conv 3×3) × 4	256×256×32	MSF-4	Maxpool (2×2,4×4,8×8) Upsample(2×2,4×4) Conv 1×1, Conv 3×3	64×64×128
DT-4	Maxpool (2×2, 4×4, 8×8) Conv 2×2, Stride 2	32×32×256	UT-1	Maxpool (2×2, 4×4, 8×8) Deconv 2×2, Stride 2	512×512×16			
E-5	(Conv 1×1, Conv 3×3) × 4	32×32×256	D-1	(Conv 1×1, Conv 3×3) × 4	512×512×16	MSF-5	Maxpool(2×2,4×4,8×8,16×16) Upsample(2×2, 4×4) Conv 1×1, Conv 3×3	32×32×256

TABLE II
ARCHITECTURAL AND OPERATIONAL DETAILS OF THE ENCODER, DECODER, AND MULTISCALE FUSION MODULES OF THE PROPOSED COVSEGNET3D FOR OPTIMUM PERFORMANCE IN INDEPENDENT SINGLE-NETWORK IMPLEMENTATION

Encoder module			Decoder module			Multi-scale fusion module		
Unit	**Ingredients**	**Output**	**Unit**	**Ingredients**	**Output**	**Unit**	**Ingredients**	**Output**
E-1	(Conv 1×1×1, Conv 3×3×3) × 2	512×512×32×16	D-4	(Conv 1×1×1, Conv 3×3×3) × 2	64×64×4×128	MSF-1	Maxpool(2×2×2,4×4×4) Upsample(2×2×2, 4×4×4,8x8x8) Conv 1×1×1, Conv 3×3×3	512×512×32×16
DT-1	Conv 2×2×2, Stride 2	256×256×16×32	UT-3	Upsample (2×2×2, 4×4×4) Deconv 2×2×2, Stride 2	128×128×8×64			
E-2	(Conv 1×1×1, Conv 3×3×3) × 2	256×256×16×32	D-3	(Conv 1×1×1, Conv 3×3×3) × 2	128×128×8×64	MSF-2	Maxpool (2×2×2, 4×4×4) Upsample(2×2×2,4×4×4) Conv 1×1×1, Conv 3×3×3	256×256×16×32
DT-2	Maxpool 2×2×2 Conv 2×2×2, Stride 2	128×128×8×64	UT-2	Maxpool (2×2×2, 4×4×4, 8×8×8) Deconv 2×2×2, Stride 2	256×256×16×32			
E-3	(Conv 1×1×1, Conv 3×3×3) × 2	128×128×8×64	D-2	(Conv 1×1×1, Conv 3×3×3) × 2	256×256×16×32	MSF-3	Maxpool (2×2×2, 4×4×4) Upsample(2×2×2,4×4×4) Conv 1×1×1, Conv 3×3×3	128×128×8×64
DT-3	Maxpool 2×2×2 Conv 2×2×2, Stride 2	64×64×4×128	UT-1	Maxpool (2×2×2, 4×4×4, 8×8×8) Deconv 2×2×2, Stride 2	512×512×32×16			
E-4	(Conv 1×1×1, Conv 3×3×3) × 2	64×64×4×128	D-1	(Conv 1×1×1, Conv 3×3×3) × 2	512×512×32×16	MSF-4	Maxpool (2×2×2,4×4×4) Upsample(2×2×2,4×4×4, 8×8×8) Conv 1×1×1, Conv 3×3×3	64×64x4×128

the network $\mathcal{N}$ can be represented as

$$\mathcal{N} = \mathbf{D_m}(\mathbf{E_m} \ldots (\mathbf{D_1}(\mathbf{E_1}(\theta_{\mathbf{E_1}}), \theta_{\mathbf{D_1}}), \ldots, \theta_{\mathbf{E_m}}), \theta_{\mathbf{D_m}}) \quad (4)$$

where $\mathbf{E_i}$ and $\mathbf{D_i}$ represent the encoder and decoder modules, respectively, of ith stage from total m stages, and θ_{E_i} and θ_{D_i} represent their respective parameters. Two-stage implementation of this architecture is schematically presented in Fig. 2.

This network can be extended from level-1 to level-L to produce a deeper variant. The encoder/decoder module constitutes of several unit cells operating at each level of the network. To generate a deeper network, additional unit cells are integrated in each of the encoder/decoder module to increase number of levels. Here, $E_{i,j}$ and $D_{i,j}$ represent the ith unit cell of jth stage of encoder and decoder, respectively, where $i = \{1, 2, \ldots, L\}$, and $j = \{1, 2, \ldots, m\}$. Hence, L number of different scales of representative feature maps are obtained from each encoder/decoder module. Moreover, scale transition of feature maps is carried out in between succeeding encoder/decoder unit cells, and effective transformation on each scale of feature maps are integrated utilizing the generalized unit cell structure in encoder/decoder module.

In between successive encoder/decoder modules, an MSF module is introduced to reduce the semantic gap with preceding stages as well as to improve the gradient propagation through parallel linkage of multiscale features. Similar to encoder/decoder module, each MSF module consists of several operational unit cells operating at different levels. Let consider $\mathbf{F_i}$ represents the ith MSF module, $F_{i,j}$ represents the ith unit cell of jth MSF module, such that $i = \{1, 2, \ldots, L\}$, $j = \{1, 2, \ldots, 2m-1\}$, and $F_{i,j} \in \mathbf{F_i}$.

Each MSF module takes all scales of feature representations as input from all preceding encoder/decoder stages, and generates L number of different feature maps for the following encoder/decoder stage through deep fusion of multiscale features obtained from preceding stages. In each unit cell of MSF module, multiscale feature aggregation and PF scheme is employed, which can be represented as

$$F_{i,j} = \mathscr{F}(\mathbf{E_1}, \mathbf{E_2}, \ldots, \mathbf{E}_{\lceil \frac{\mathbf{j}}{\mathbf{2}} \rceil}, \mathbf{D_1}, \mathbf{D_2}, \ldots, \mathbf{D}_{\lfloor \frac{\mathbf{j}}{\mathbf{2}} \rfloor})$$
$$\forall\, i = \{1, 2, \ldots, L\},\ j = \{1, 2, \ldots, 2m-1\} \quad (5)$$

where $\mathscr{F}(.)$ represents the functional operations in the MSF unit cell considering L scale of representations from each of the preceding encoder/decoder module.

From final level of the sequential decoder modules, several decoded feature representations are obtained, which are processed together in the fusion optimizer unit ($\mathcal{O}$) to produce the final segmentation mask, and it can be given by

$$\mathcal{O} = \mathcal{F}(D_{1,1}, D_{1,2}, \ldots, D_{1,m}) \quad (6)$$

where $\mathcal{O}(.)$ represents the fusion optimizer function.

All the basic building blocks of the CovSegNet architecture are generic and can be designed and optimized for both 2-D and 3-D operations. In the following discussions, different building blocks of the CovSegNet architecture are presented in detail. For the ease of discussion, mainly 2-D operational blocks are focused. However, for 3-D operations, all the convolutional kernels, pooling/upsampling windows are shifted in dimension for operating with 3-D voxels instead of 2-D pixels. Architectural details for the most optimized implementations of CovSegNet2D and CovSegNet3D are presented in Tables I and II, respectively.

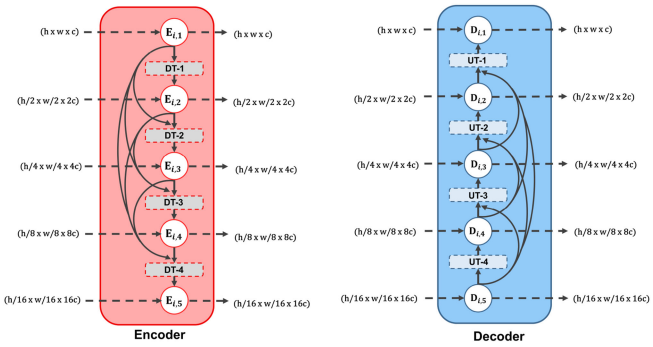

Fig. 3. Schematic representations of the proposed encoder and decoder modules in five-level implementation having five unit blocks along with associated down transition (DT)/up transition (UT) units in between subsequent unit blocks. Here, (h, w, c) is used to denote the height, width, and channel of the feature maps at different phase. (a) Encoder module. (b) Decoder module.

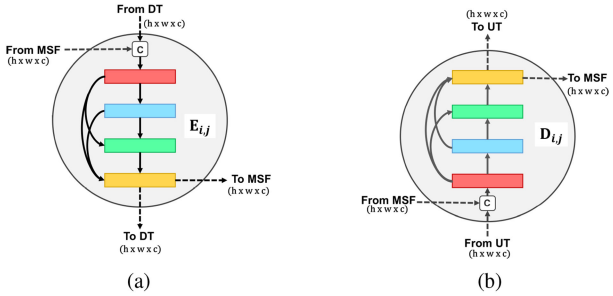

Fig. 4. Structure of the encoder/decoder unit cells. Four densely interconnected convolutional stages are employed in each unit. Here, "c" denotes the channelwise concatenation of feature maps extracted from transition unit and MSF unit. $E_{i,j}/D_{i,j}$ denote the unit blocks of ith level in jth module. (a) Encoder unit cell. (b) Decoder unit cell.

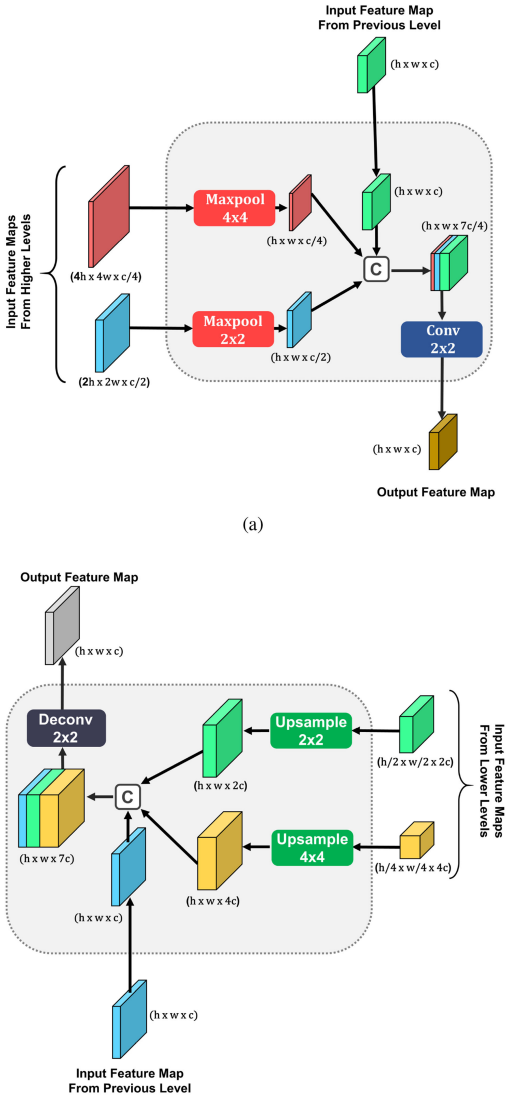

Fig. 5. Schematic representations of the DT unit (operating between level-3 and -4) and the UT unit [operating between level-$(L-2)$ and level-$(L-3)$]. All the feature maps generated from preceding unit blocks are made uniform and integrated in the transition process. (a) DT unit (DT-3). (b) UT unit (UT-2).

C. Proposed Encoder/Decoder Structure

The encoder and decoder modules are structurally similar that are successively used in the sequential stages of CovSegNet. Encoder/decoder modules are schematically presented in Fig. 3. These encoder/decoder modules are composed of several operational unit cells with transitional dense interconnections. The operations of encoder/decoder modules can be divided into two categories: unit cell operations and transitional operation.

1) Encoder/Decoder Unit Cell Operation: In Fig. 4, the unit cell structure of the encoder/decoder module is presented. In each unit cell, two input feature map is entered, one from the transitional unit and the other from the preceding MSF unit, whereas the output feature map is passed through following transitional and MSF operations. Moreover, each unit cell consists of four densely interconnected convolutional layers, where each convolutional layer provides two sequential convolutional filtering with (1×1) and (3×3) kernels. Such dense interconnection between convolutional operations has been proven to be effective in numerous applications. No dimensional scaling has been carried out in each of this unit cell as it is employed for introducing adequate transformation in the feature space to encode/decode effective representation. Hence, this unit cell operations can be functionally represented as $E, D : \mathbb{R}^{h\times w\times c} \rightarrow \mathbb{R}^{h\times w\times c}$, where (h, w, c) represents the height, width, and channel of the feature map.

2) Encoder Down-Transitional Operation: During down-transitional operations between subsequent unit cells of the encoder module, the spatial dimension of the feature map is reduced for generalizing the feature map, whereas the channel depth is increased to incorporate more filtering operations in subsequent levels for generating more sparser features. It can be functionally presented as $f : \mathbb{R}^{h\times w\times c} \rightarrow \mathbb{R}^{h/2\times w/2\times 2c}$, where spatial resolution is downscaled by 2 and channel depth is increased by 2 from the input feature map obtained from the previous level. However, traditional downsampling operations using pooling/strided convolutions results in loss of contextual information. Moreover, it can be more prominent while

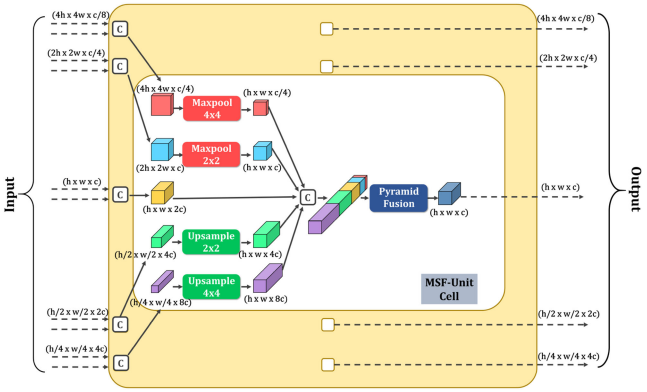

Fig. 6. Schematic representation of the proposed multiscale fusion module. Detailed operations performed in an MSF unit cell are particularly focused, and similar operations are carried out in other unit cells of the MSF module.

incorporating a deep stack of unit cells in the encoder module. To mitigate the loss of contextual information in down-transitional operation, a higher level of dense interconnection is proposed among multiscale feature maps generated from different unit cells. In Fig. 5(a), the structure of such a DT unit is schematically presented. In each of such DT unit, encoded feature representations generated from all higher levels of unit cells are considered for generating the down-scaled feature map. Hence, contextual information lost in each transitional operation can be recovered from very deep stack of unit cells as feature representations from all preceding cells are considered during transition. To converge multiscale feature maps from preceding levels, first, pooling operations with different kernels are carried out to make their spatial dimension uniform and subsequently, channelwise feature aggregation is carried out. The aggregated feature map, $F_{\text{agg,DT}}$, generated at ith level can be represented as

$$F^i_{\text{agg,DT}} = E^i \oplus P^{(2\times 2)}(E^{i-1}) \cdots \oplus P^{(2^{i-1}\times 2^{i-1})}(E^1) \quad (7)$$

where $\oplus$ indicates the feature concatenation, $P^{(2\times 2)}$ represents pooling operation with (2×2) window, and E^i represents the output of ith unit cell of the encoder.

Finally, a convolutional operation with (2×2) kernel is carried out with a stride of (2×2) for generating the downscaled feature map by filtering the aggregated feature vector.

3) Decoder Up-Transitional Operation: On the contrary, up transitional operations are carried out in between successive decoder unit cells to provide the dimensional shifting toward the reconstruction of the final segmentation mask. In each of such UT operations, spatial resolution is upscaled by 2, whereas channel depth is reduced by 2 to get closer to the final reconstruction mask and it can be represented as $f' : \mathbb{R}^{h\times w\times c} \rightarrow \mathbb{R}^{2h\times 2\, w\times c/2}$. Similar to the down-transitional operation in encoder, all the preceding representations of multiscale decoded feature maps generated from different unit cells are taken into consideration in the UT operation to gather more contextual information [see Fig. 5(b)]. First, spatially uniform feature maps are created through bilinear interpolation upsampling with different windows, and feature aggregation is carried out to generate

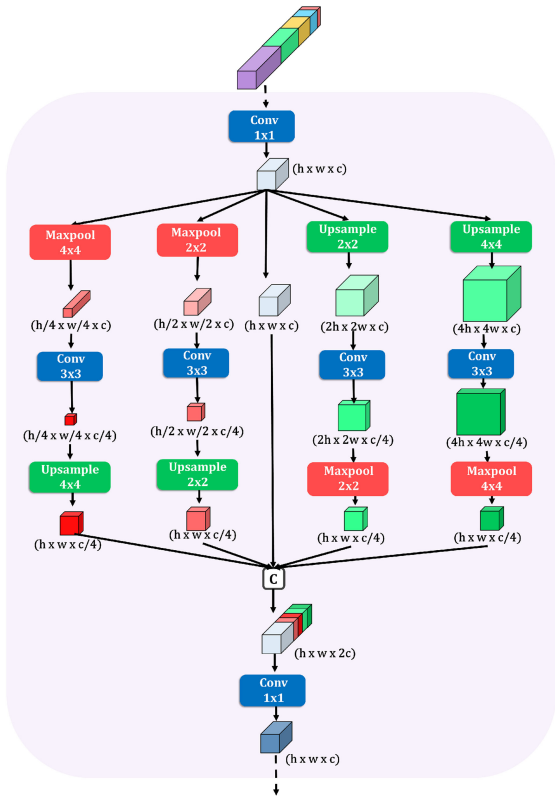

Fig. 7. Proposed PF scheme utilizing diverse windows of frequent upsampling and downsampling operations for fusing multiscale features.

aggregated feature vector $F_{\text{agg,UT}}$, which is given by

$$F^i_{\text{agg,UT}} = D^i \oplus U^{(2\times 2)}(D^{i+1}) \cdots \oplus U^{(2^{i-1}\times 2^{i-1})}(D^L) \quad (8)$$

where $U^{(2\times 2)}$ represents bilinear upsampling operation with (2×2) window and D^i represents the output of ith unit cell of the decoder.

Finally, the aggregated feature map is processed using a deconvolution operation with (2×2) kernel to incorporate the necessary dimensional up-scaling for further processing in the following unit cell.

D. Proposed MSF Module With PF Scheme

During sequential encoding–decoding operations, a semantic gap is generated between a similar scale of encoded and decoded feature maps. Moreover, in the traditional architecture, the gradient has to propagate sequentially that sometimes gives rise to vanishing gradient problems for deeper encoder/decoder module particularly. As multiple stages of encoding and decoding operations are integrated into the CovSegNet, this problem is supposed to be more prominent if all the encoder and decoder modules are sequentially connected. To overcome these limitations, an MSF module is proposed that develops parallel interconnection among different scales of feature maps of the encoder/decoder modules utilizing a PF scheme.

As shown in Fig. 6, each MSF module consists of several MSF unit cells where each cell considers multiscale feature maps

generated from different levels of preceding encoder/decoder modules and generates feature map for the unit cell of the following encoder/decoder module. Here, similar scale of feature representations generated from different levels of the preceding encoder/decoder modules are concatenated, first, to produce L number of multiscale feature maps. Afterward, all the L scales of feature maps are made spatially equivalent in dimension through pooling and bilinear upsampling with different windows, and channelwise feature concatenation is carried out to generate the aggregated feature vector. This can be represented as

$$F_{\text{agg,MSF}}^{(i,j)} = P^{(2^{i-1}\times 2^{i-1})}(f_1) \oplus \cdots \oplus P^{(2\times 2)}(f_{i-1}) \oplus \oplus f_i \oplus U^{(2\times 2)}(f_{i+1}) \oplus \cdots \oplus U^{(2^{L-i}\times 2^{L-i})}(f_L) \quad (9)$$

$$f_i = E_{(1,j)} \oplus \ldots E_{(i,j)} \oplus D_{(1,j)} \oplus \cdots \oplus D_{(i-1,j)} \quad (10)$$

where $F_{\text{agg,MSF}}^{(i,j)}$ is the aggregated feature vector generated in the ith level of jth MSF module, and f_i represents the ith concatenated feature map.

Afterward, the aggregated feature vector is passed through a PF scheme to generate the output feature vector that will be fed to the corresponding encoder/decoder unit cell of the following module. Hence, the generated output feature map from each MSF unit cell contains information from all preceding modules and thus, establishes a parallel flow of optimization for efficient gradient propagation.

E. Proposed PF Module

The PF module incorporates PF scheme into the aggregated feature map of MSF unit cell ($F_{\text{agg,MSF}}$) utilizing the combinations of sequential multiwindow pooling and upsampling operations (see Fig. 7). First, the depth of the aggregated vector $F_{\text{agg,MSF}}$ is reduced through a pointwise convolution (kernel, 1×1) to generate feature vector f_a, and thus, $F_{\text{agg,MSF}} \mapsto f_a$, where $f_a \in \mathbb{R}^{h\times w\times c}$.

Afterward, the generated vector f_a passes through multiple spatial scaling-vertical scaling-inverse spatial scaling operations in parallel with different scaling factors. Spatial scaling operation is carried out utilizing pair of pooling and upsampling operations with different kernel windows, whereas vertical scaling is employed utilizing convolutional filtering (kernel, 3×3) to reduce the channel depth by one-fourth of the initial depth. Initial reduction followed by expansion of the feature map assists in gathering the more general feature representation, whereas initial expansion followed by reduction of the feature map gathers the more detailed information from a sparser domain. These operations pave the way to extract the most generalized representations through analyzing from diverse feature domains, which can be represented by

$$P_r : \mathbb{R}^{h\times w\times c} \to \mathbb{R}^{h*r\times w*r\times c} \to \mathbb{R}^{h*r\times w*r\times c/4} \to \mathbb{R}^{h\times w\times c/4} \quad \forall r = \{0.25, 0.5, 2, 4\} \quad (11)$$

where P_r denotes one of the parallel operational paths in the PF module with a spatial scaling factor of r.

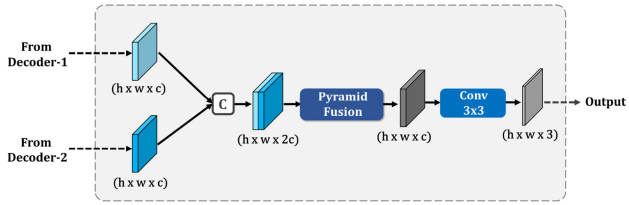

Fig. 8. Schematic of the fusion optimizer module optimizing the decoded feature maps generated from two decoding stages.

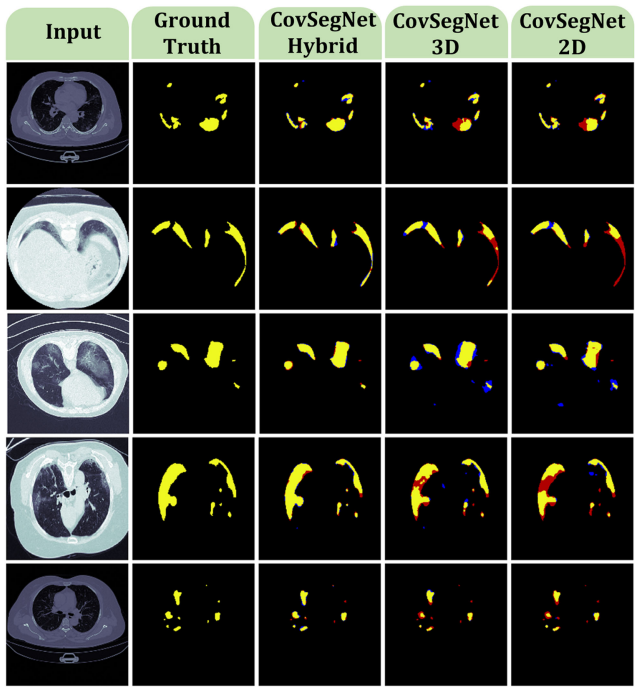

Fig. 9. Visual representations of the segmentation performances obtained using single-phase training (CovSegNet2D and CovSegNet3D) and multiphase training (with hybrid 2-D–3-D networks) in Dataset-1. Here, "yellow" represents the TP regions, "red" represents the FN regions, and "blue" represents the FP regions.

Afterward, feature aggregation operation is carried out utilizing different representations generated at multiple paths along with the input representation to generate the aggregated vector $F_{\text{agg,PF}}$, where $F_{\text{agg,PF}} \in \mathbb{R}^{h\times w\times 2c}$. Finally, a final pointwise convolution (kernel, 1×1) is carried out to generate the output feature map $f_{\text{out,PF}}$, where $f_{\text{out,PF}} \in \mathbb{R}^{h\times w\times c}$.

F. Structure of the Fusion Optimizer($\mathcal{O}$)

The decoded feature maps generated from the top of decoder modules are considered for final reconstruction through a fusion optimization process. This process is schematically shown in Fig. 8. Initially, an aggregated feature vector $F_{\text{agg},\mathcal{O}}$ is created considering all the output feature maps from different decoder modules, which can be given by

$$F_{\text{agg},\mathcal{O}} = D_{1,1} \oplus D_{1,2} \oplus \cdots \oplus D_{1,S} \quad (12)$$

where S denotes the total number of stages.

TABLE III
ABLATION STUDY OF THE EFFECT OF DIFFERENT MODULES IN THE PERFORMANCE (MEAN ± STANDARD DEVIATION) OF THE PROPOSED COVSEGNET2D ARCHITECTURE

Network	Dataset-1					Dataset-2				
	Sen.(%)	Spec.(%)	Dice(%)	IoU(%)	p-Value	Sen.(%)	Spec.(%)	Dice(%)	IoU(%)	p-Value
Baseline (V1)	82.7± 0.49	97.4± 0.09	84.1±0.29	79.8±0.21	-	71.7±0.12	95.8±0.18	71.9±0.33	65.8±0.27	-
Baseline+ DT (V2)	83.8±0.29	97.8±0.12	85.8±0.36	81.1±0.08	0.0033	73.6±0.31	96.5±0.15	73.4±0.14	67.6±0.21	0.0023
Baseline+ UT (V3)	83.1±0.25	97.7±0.08	85.4±0.16	80.9±0.13	0.0017	73.1±0.55	96.3±0.18	73.1±0.19	67.2±0.35	0.0044
Baseline+ DT+UT (V4)	84.9±0.41	98.1±0.11	86.7±0.27	82.3±0.32	0.0021	74.6±0.17	97.1±0.12	74.8±0.34	69.4±0.18	0.0012
Baseline+(MSF-w/o PF) (V5)	86.9±0.15	98.3±0.07	87.3±0.28	82.9±0.26	0.0019	76.2±0.27	97.9±0.16	77.2±0.29	72.8±0.24	0.0034
Baseline+ MSF (V6)	88.4±0.28	98.7±0.08	89.2±0.32	84.1±0.21	0.0041	78.8±0.25	98.4±0.11	79.5±0.21	74.1±0.25	0.0048
CovSegNet2D (V7)	90.8±0.32	99.1±0.13	91.1±0.25	86.9±0.09	0.0011	81.5±0.22	98.9±0.13	82.7±0.08	77.5±0.14	0.0009

TABLE IV
ABLATION STUDY OF THE EFFECT OF DIFFERENT MODULES IN THE PERFORMANCE (MEAN ± STANDARD DEVIATION) OF THE PROPOSED COVSEGNET3D ARCHITECTURE IN DATASET-1

Network	Dataset-1				
	Sensitivity(%)	Specificity(%)	Dice Score(%)	IoU(%)	p-Value
Baseline3D (V1$_{3D}$)	84.5±0.21	97.9±0.12	85.2±0.23	80.8±0.32	-
Baseline3D + DT (V2$_{3D}$)	85.7±0.31	98.2±0.19	86.1±0.25	82.3±0.29	0.0011
Baseline3D + UT (V3$_{3D}$)	85.2±0.18	98.1±0.08	85.9±0.18	82.0±0.21	0.0008
Baseline3D + DT+UT (V4$_{3D}$)	86.7±0.22	98.7±0.14	88.3±0.28	83.5±0.27	0.0017
Baseline3D+(MSF-w/o PF) (V5$_{3D}$)	87.4±0.25	97.9±0.11	88.2±0.21	83.8±0.31	0.0032
Baseline3D+ MSF (V6$_{3D}$)	89.6±0.19	98.4±0.15	89.9±0.17	85.1±0.19	0.0021
CovSegNet3D	91.1±0.26	99.3±0.09	92.3±0.15	87.7±0.23	0.0025

Afterward, PF scheme is employed on aggregated vector to obtain the more generalized representation utilizing multiscale decoded representations. Finally, another convolutional filtering (kernel, 3×3) is carried out to generate the final segmentation mask f_{mask}, utilizing binary activation function, and these can be represented as

$$f_{\text{mask}} = \sigma(\text{Conv}(\text{PF}(F_{\text{agg},\mathcal{O}})) \tag{13}$$

where $\sigma(.)$ denotes the nonlinear activation.

G. Loss Function

Tversky index is introduced in [31] for better generalization of the dice index by balancing out FPs and FNs, which is given by

$$\text{TI} = \frac{\sum_{i=1}^{P} p_{1i} g_{1i} + \epsilon}{\sum_{i=1}^{P} p_{1i} g_{1i} + \alpha \sum_{i=1}^{P} p_{0i} g_{1i} + \beta \sum_{i=1}^{P} p_{1i} g_{0i} + \epsilon} \tag{14}$$

where g_{0i} and p_{0i} indicate, respectively, the ground truth and prediction probability of pixel i being in a normal region, whereas g_{1i} and p_{1i} indicate, respectively, the ground truth and prediction probability of pixel i being in an abnormal region, P is the total number of pixels on a certain image, α and β are used to shift emphasize for balancing class imbalance such that $\alpha + \beta = 1$, and $\epsilon(10^{-8})$ is used to avoid division-by-zero as safety factor.

To put more emphasis on hard training examples, a focal Tversky loss function is introduced in [32] utilizing the Tversky index, which is given by

$$\mathcal{L} = \sum_{c} (1 - TI_c)^{\frac{1}{\gamma}} \tag{15}$$

where γ is used to emphasize the challenging less accurate predictions. Due to the better generalization over a large number of datasets according to Abraham and Khan [32], $\alpha = 0.7, \beta = 0.3$, and $\gamma = \frac{4}{3}$ are used for all experimentations in this article.

If $\mathbf{y}$ and $\mathbf{y^P}$ denote slicewise mask ground truth and corresponding probability prediction, respectively, whereas

TABLE V
EFFECT OF VERTICAL EXPANSIONS (LEVELS) AND HORIZONTAL EXPANSIONS (STAGES) ON THE DICE SCORE (MEAN ± STANDARD DEVIATION) IN DATASET-1

Level	CovSegNet2D			CovSegNet3D		
	One-stage	Two-stage	Three-stage	One-stage	Two-stage	Three-stage
2	49.9±0.37	75.3±0.13	78.12±0.21	57.3±0.18	79.8±0.18	82.1±0.19
3	64.8±0.23	85.8±0.32	88.5±0.15	69.3±0.35	89.2±0.26	90.2±0.25
4	75.2±0.32	89.6±0.27	90.8±0.22	79.8±0.29	92.3±0.15	91.8±0.17
5	83.5±0.19	91.1±0.25	89.9±0.12	84.5±0.43	90.2±0.34	89.7±0.28
6	86.7±0.27	90.9±0.21	89.1±0.11	89.3±0.21	89.8±0.41	87.9±0.36

$\mathbf{Y}$ and $\mathbf{Y^P}$ denote volumetric mask ground truth and corresponding probability prediction, respectively, the objective loss functions for separately optimizing CovSegNet2D and CovSegNet3D can be represented as

$$\mathscr{L}_{\text{2D}} = \mathcal{L}(\mathbf{y}, \mathbf{y^P});\ \mathbf{y}, \mathbf{y^P} \in \mathbb{R}^{h \times w \times c} \tag{16}$$

$$\mathscr{L}_{\text{3D}} = \mathcal{L}(\mathbf{Y}, \mathbf{Y^P});\ \mathbf{Y}, \mathbf{Y^P} \in \mathbb{R}^{h \times w \times s \times c}. \tag{17}$$

The joint optimization objective function used in phase-2 combining slicewise and volumetric operations is given by

$$\mathcal{F} = \lambda \left(\frac{1}{S} \sum_{i=1}^{S} \mathscr{L}_{\text{2D}}^{i} \right) + \mathscr{L}_{\text{3D}} \tag{18}$$

where λ denotes the scaling factor of 2-D loss term, and s denotes total number of 2-D slices per volume. Here, $\lambda = 0.2$ is used for optimization to provide more emphasis on CovSegNet3D in phase-2 as CovSegNet2D is pretrained in phase-1 and is supposed to be fine-tuned in phase-2.

III. RESULTS AND DISCUSSIONS

Experimentations have been carried out on three publicly available datasets to validate the effectiveness of the proposed scheme on numerous segmentation tasks. Performances of CovSegNet2D and CovSegNet3D have been separately studied along with the proposed hybrid scheme of joint optimization combining CovSegNet2D and CovSegNet3D.

TABLE VI
PERFORMANCE COMPARISON (MEAN ± STANDARD DEVIATION) OF THE PROPOSED COVSEGNET2D ARCHITECTURE WITH OTHER STATE-OF-THE-ART APPROACHES ON 2D-CT SLICES

Network	Dataset-1					Dataset-2				
	Sen.(%)	Spec.(%)	Dice(%)	IoU(%)	p-Value	Sen.(%)	Spec.(%)	Dice(%)	IoU(%)	p-Value
Unet [22]	75.9±0.34	88.9±0.12	79.3±0.26	74.9±0.18	-	52.9±0.29	86.2±0.09	43.3±0.34	38.8±0.32	-
Unet++ [28]	78.6±0.17	91.1±0.18	81.1±0.23	76.2±0.21	-	57.7±0.32	89.2±0.11	52.3±0.31	48.1±0.37	-
MultiResUnet [27]	77.2±0.33	90.3±0.24	82.7±0.28	77.4±0.15	-	56.9±0.27	86.9±0.15	50.8±0.28	45.2±0.22	-
Attention-Unet-2D [29]	81.1±0.29	92.2±0.11	85.1±0.14	79.6±0.28	-	60.8±0.25	88.4±0.12	57.7±0.36	51.9±0.26	-
CPF-Net [30]	78.9±0.27	91.7±0.14	84.4±0.25	79.3±0.25	-	62.2±0.14	91.1±0.14	60.4±0.25	56.1±0.21	-
Semi-Inf-Net [12]	82.7±0.26	94.8±0.21	86.9±0.34	81.1±0.18	-	72.9±0.44	95.8±0.19	74.1±0.24	68.1±0.32	-
CovSegNet2D(Ours)	90.8±0.32	99.1±0.13	91.1±0.25	86.9±0.09	0.0008	81.5±0.22	98.9±0.13	82.7±0.08	77.5±0.14	0.0013

Fig. 10. Visual representations of the segmentation performances of different state-of-the-art networks on the CT images from Database-1 and Database-2. Here, "yellow" represents the true positive (TP) regions, "red" represents the false negative (FN) regions, and "blue" represents the false positive (FP) regions.

TABLE VII
PERFORMANCE COMPARISON (MEAN ± STANDARD DEVIATION) OF THE COVSEGNET3D ARCHITECTURE WITH OTHER STATE-OF-THE-ART NETWORKS ON 3-D CT VOLUMES OF DATASET-1

Network	Sen.(%)	Spec.(%)	Dice(%)	IoU(%)	p-Value
Unet-3D [22]	77.1±0.22	89.8±0.18	84.2±0.27	79.4±0.24	-
Unet++-3D [28]	79.2±0.17	91.7±0.25	85.1±0.29	80.2±0.26	-
MultiResUnet-3D [27]	78.7±0.27	90.9±0.16	84.5±0.31	78.9±0.18	-
Attention-Unet-3D [29]	82.5±0.26	93.1±0.31	85.9±0.24	81.4±0.29	-
CPF-Net-3D [30]	80.1±0.23	92.6±0.23	85.2±0.18	80.8±0.34	-
VNet-3D [19]	84.3±0.29	93.9±0.17	85.7±0.31	81.3±0.19	-
CovSegNet3D(Ours)	91.1±0.26	99.3±0.09	92.3±0.15	87.7±0.23	0.0024
CovSegNet-Hybrid(Ours)	92.6±0.25	99.5±0.07	94.1±0.19	90.2±0.27	0.0011

A. Dataset Description

Dataset-1 contains 20 CT volumes with 1800+ slices annotated by expert radiologist panel [33]. All the slices have annotations for both lung and infection regions. Each slices are of resolution (630×630), which are resized to (512×512). Dataset-2 is the "COVID-19 CT Segmentation dataset" that contains 110 axial CT images collected by the Italian Society of Medical and Interventional Radiology from 40 different COVID patients [34]. All the images are of resolution (512×512). Each slice contains multiclass annotations of infections. Dataset-3 is the "Semantic Drone Dataset" where the semantic understanding of urban scenes is mainly focused to increase the safety of drone flight and landing procedures [35]. This dataset consists of 400 images with pixelwise annotation for 20 different classes having resolutions of 6000×4000 and all of these images are resized to (512×512). Experimentations on Dataset-3 is mainly integrated to investigate the effectiveness of the proposed CovSegNet architecture on other domains with challenging operating conditions.

TABLE VIII
EFFECT OF DIFFERENT LOSS FUNCTIONS ON THE PERFORMANCE [DICE SCORE(%)] OF COVSEGNET ON DATASET-1

Loss function	CovSegNet2D	CovSegNet3D	CovSegNet-Hybrid
IoU Loss	89.9±0.23	90.7±0.16	92.4±0.12
Dice Loss	90.2±0.13	91.1±0.21	93.3±0.09
Dice Loss+ BCE loss	90.4±0.11	91.5±0.17	93.6±0.15
Focal Tversky loss	91.1±0.25	92.3±0.15	94.1±0.19

B. Experimental Setup

Different hyperparameters of the network are chosen through experimentation for better performance. Adam optimizer is employed for optimization of the network during the training phase with an initial learning rate of 10^{-5}. The learning rate is decayed after ever ten epochs with a decaying rate of 0.99. Intel Xeon $D-1653$ N CPU @2.80 GHz with 12 M Cache and 8 cores along with 24-GB RAM is used for experimentation. For hardware acceleration, $2\times$ NVIDIA RTX 2080 Ti GPU having with 4608 CUDA cores running 1770 MHz with 24-GB GDDR6 memory is deployed. The network is trained for 1000 epochs on each dataset. Batch size is chosen to be 32 for processing 2-D CT slices, whereas it is chosen to be 2 for processing 3-D CT volume.

A number of traditional evaluation metrics are used for the evaluation of performance. These are given by

$$\text{IoU} = \frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}} \tag{19}$$

$$\text{Dice Score} = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}} \tag{20}$$

$$\text{Specificity} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{21}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{22}$$

where TP, FP, and FN denote true positive, false positive, and false negative predictions, respectively. A five-fold cross-validation scheme is carried out separately on these databases for evaluation of the proposed scheme. Mean and standard deviations of the evaluation metrics obtained from different test folds are reported. For binary thresholding of the predicted probability mask, a threshold of 0.5 is used in general. The Wilcoxon rank-sum test is used for statistical analysis of the performance improvement obtained from the proposed scheme. The performances of the proposed schemes are statistically analyzed and the statistical significance level is set to $\alpha = 0.01$. The null hypothesis is that no significant improvement of performance is achieved using the proposed scheme over the other existing best performing approaches.

C. Ablation Study

To analyze the effectiveness of different modules of the proposed CovSegNet architecture, an ablation study is carried out. The baseline model is defined as the two-stage implementations with encoder and decoder modules only excluding the DT units, UT units, and MSF modules. The statistical significance test is carried out to validate the improvement of dice scores over the baseline model.

1) Effects of the Transition Unit: Instead of proposed DT units and UT units, traditional max-pooling and upsampling operations are used, respectively, in the baseline model according to the conventions of the traditional Unet architecture. Performances with different combinations of transition units are provided in (V2–V4) of Table III for 2-D analysis. The inclusion of DT unit (V2) in encoder modules provides 1.7% improvement and 1.5% improvement of dice scores in Database-1 and 2, respectively, over the baseline. Moreover, the inclusion of UT unit (V3) in decoder modules provides 1.3% and 1.2% improvements of dice scores, whereas the inclusion of both of the transition units (V4) provide 2.6% and 2.9% improvements of dice scores in Database-1 and -2, respectively. Hence, both of the UT units and DT units are contributing considerable improvements over the baseline performance. Similar improvements can be noticeable for 3-D variants of the transition units also (from $V2_{\text{3D}}$ to $V4_{\text{3D}}$) that are summarized in Table IV. All the improvements are found to be statistically significant ($p < 0.01$).

2) Effects of the MSF Module: The MSF modules are proposed in place of the traditional direct skip connection scheme of Unet architecture to reduce the semantic gaps between subsequent encoder and decoder modules. In the baseline model, direct skip connections are used between succeeding modules instead of the MSF module. In Table III, the change of performance with the inclusion of the MSF module in the 2-D baseline model is provided in V6. It should be noticed that 5.1% improvement of dice score and 4.3% improvement of IoU score have been achieved in Database-1, whereas 7.6% improvement of dice score and 8.3% improvement of IoU score have been achieved in Database-2. Similar performance improvements can be noticed for the incorporation of MSF module in the 3-D baseline model ($V6_{\text{3D}}$ in Table IV). These improvements are found to be statistically significant ($p < 0.01$).

3) Effects of the PF Scheme in MSF Module: PF modules are integrated into the MSF modules to operate on the aggregated multiscale feature vector in the MSF module. Instead of the PF module, a pointwise convolution with (1×1) kernel can be performed to reduce and transform the aggregated vector into the output vector. The performance of the 2-D baseline model, including this simplified version of the MSF module, is reported in V5 of Table III. It is to be noted that 2.3% improvement of dice score is achieved in Database-1 and 3.4% improvement is achieved in Database-2 over the baseline model using these simplified MSF modules, and these improvements are statistically significant ($p < 0.01$). However, 3.2% and 5.3% reduction of dice scores can be noticed in Database-1 and -2, respectively, from the baseline model with original MSF modules (V6) incorporating PF scheme. Similarly, considerable improvement is also achieved for the incorporation of 3-D PF scheme in the 3-D variants of MSF module, which can be noticed from $V5_{\text{3D}}$ and $V6_{\text{3D}}$ in Table IV. It justifies the effectiveness of the PF scheme in the MSF module.

4) Effects of Vertical and Horizontal Scaling: The proposed CovSegNet architecture is designed in a modular way with the opportunity for both vertical and horizontal expansions for

TABLE IX
COMPARISON OF PERFORMANCES (MEAN ± STANDARD DEVIATION) ON DIFFERENT TYPES OF INFECTIONS (GGO AND CONSOLIDATION) IN DIFFERENT CT-SLICES OF DATASET-2

Network	Consolidation					GGO				
	Sen.(%)	Spec.(%)	Dice(%)	IoU(%)	p-Value	Sen.(%)	Spec.(%)	Dice(%)	IoU(%)	p-Value
Unet [22]	41.1±0.26	96.2±0.12	40.3±0.28	35.5±0.28	-	35.1±0.27	98.2±0.09	44.1±0.27	39.8±0.25	-
Unet++ [28]	48.8±0.23	97.8±0.16	42.6±0.26	38.2±0.19	-	41.2±0.32	96.6±0.14	49.9±0.22	45.7±0.27	-
MultiResUnet [27]	46.6±0.28	97.1±0.14	42.1±0.19	37.6±0.27	-	44.5±0.28	97.3±0.11	47.7±0.18	43.1±0.28	-
Attention-Unet-2D [29]	44.8±0.19	96.8±0.08	44.5±0.25	40.1±0.33	-	55.3±0.31	95.4±0.08	52.9±0.17	47.6±0.35	-
CPF-Net [30]	49.9±0.18	97.4±0.15	44.1±0.23	39.9±0.29	-	53.5±0.22	96.9±0.13	56.9±0.26	51.1±0.34	-
Semi-Inf-Net [12]	50.9±0.22	96.7±0.11	45.8±0.31	41.4±0.18	-	62.2±0.34	96.1±0.18	62.7±0.22	58.4±0.23	-
CovSegNet2D(Ours)	63.8±0.17	98.4±0.09	56.8±0.24	51.9±0.25	0.0017	73.3±0.25	98.9±0.12	70.9±0.31	66.1±0.19	0.0028

TABLE X
COMPARISON OF PERFORMANCES (MEAN ± STANDARD DEVIATION) ON MULTICLASS SEMANTIC SEGMENTATION TASK OF DATASET-3

Network	Dataset-3				
	Sen.(%)	Spec.(%)	Dice(%)	IoU(%)	p-Value
Unet [22]	56.9±0.19	68.6±0.23	42.2±0.35	37.7±0.28	-
Unet++ [28]	57.3±0.25	70.4±0.31	44.8±0.29	40.1±0.33	-
Attention-Unet-2D [29]	58.7±0.22	71.8±0.29	48.5+0.42	43.9±0.25	-
CPF-Net [30]	61.5±0.28	73.1±0.17	51.4±0.38	47.7±0.34	-
Semi-Inf-Net [12]	64.9±0.31	76.3±0.27	50.9±0.27	46.4±0.26	-
CovSegNet(ours)	76.4±0.18	87.7±0.16	64.6±0.21	59.5±0.29	6e-5

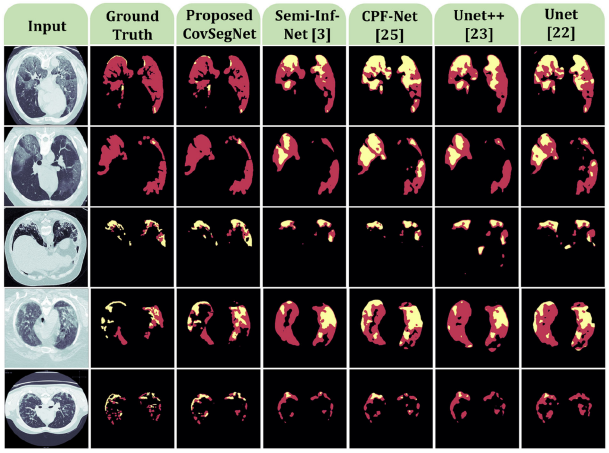

Fig. 11. Visual representations of the segmented multiclass lesions of the CT images from Database-2 obtained using different state-of-the-art networks. Here, "red" represents the "GGO" regions and "yellow" represents the "consolidation" regions.

integrating more number of levels and stages, respectively. In Table V, the performances of the CovSegNet architecture with different numbers of levels and stages are provided. It should be noticed that the optimum dice score of 91.1% is obtained for CovSegNet2D with five levels and two stages. The best performance on single-stage implementation is found to be 86.7%, which is 4.4% lower than the best of the two-stage implementation. Similar analyses have been carried out on CovSegNet3D using volumetric data where the highest dice score of 92.3% is achieved with three-levels and two-stages implementation.Moreover, when more stages are included, comparably higher performances are obtained in a lower number of levels, e.g., best dice score of 90.8% in the three-stage setup of CovSegNet2D has been achieved with four levels. With the horizontal expansion, the model gathers more amount of contextual information in a lower number of stages that result in higher performances. However, more expansion in both directions starts to increase the complexity that causes a decrease in performance due to overfitting issues.

5) Effects of the Hybrid 2-D–3-D Joint Optimization Scheme With Two-Phase Training: The proposed two-phase training scheme exploits the advantages of both the slice-based optimization and volumetric optimization. Quantitative performances obtained using CovSegNet2D, CovSegNet3D, and the hybrid scheme are provided in Tables VI and VII. Slice-based processing provides the advantages of employing deeper networks for lighter 2-D convolutions, whereas loses the interslice contextual information that results in suboptimal performance. On the other hand, 3-D volumetric analysis incorporates more contextual information while increasing the computational burden of optimization for the expensive 3-D kernels processing. The best variant of CovSegNet3D provides 1.2% higher dice score, and 0.8% higher IoU score over the best variant of CovSegNet2D. Thus, the performances of the proposed CovSegNet architectures are quite comparable in both 2-D and 3-D processing with minor variations. It is to be noted that more improvements can be achieved with the expensive 3-D processing if the number of training CT volumes can be increased substantially for exploiting the advantages of the complete 3-D processing. However, by combining the advantages of both these schemes in the proposed multiphase hybrid training approach, 3% and 1.8% higher dice scores are achieved compared to the best performing CovSegNet2D and CovSegNet3D architectures, respectively. In the hybrid scheme, to reduce the computational burden of 3-D data processing, only two-level and dual-stage implementation of the CovSegNet3D is employed accompanied by the four-level and dual-stage implementation of the CovSegNet2D that provides the optimal performance with minimal complexity. Since a very shallower variant of CovSegNet3D is employed in the hybrid network compared to the best performing variant of CovSegNet3D, the operational complexity is greatly reduced in the hybrid network that led to the optimum performance with the available CT volumes. This improvement signifies the effectiveness of the hybrid networking scheme in multiphase training ($p < 0.01$). Moreover, qualitative analysis of the performances of the individual networks and hybrid networks are presented in Fig. 9 with different levels of infection. It should be noticed that both of the FP and FN regions are reduced in the segmented mask for the hybrid scheme compared to the individual networks. Therefore, for the proper optimization with the hybrid networking scheme through multiphase training, optimum performance is achieved compared to the independent 2-D/3-D data processing.

6) Effects of the Loss Functions: In Table VIII, effects of different loss functions are summarized on the performance of the CovSegNet. For optimizing the hybrid network, joint optimization objective function [see (18)] is defined incorporating losses of the CovSegNet2D and CovSegNet3D networks. Several traditional loss functions are experimented to evaluate the effects of loss functions on the performance of the proposed network. It should be noticed that focal Tversky loss function provides 0.9% improvement of dice score over traditional dice loss function, 1.7% improvement over IoU loss, and 0.7% improvement over the aggregated dice loss and binary cross entropy loss function. Despite the slight variations of performance with different loss functions, it is to be noted that the proposed CovSegNet-hybrid network consistently provides considerably better performance over other traditional networks with any of these loss functions. Since the available contextual information are effectively exploited through the proposed hybrid learning scheme along with numerous architectural renovations, the proposed network shows very stable and comparable performance with different loss functions. Such phenomenon signifies the robustness of the proposed scheme for extracting the effective feature from 3-D CT volumes to achieve optimum performance irrespective of the loss functions.

D. Comparison With Other Existing Approaches

To compare the performances of the proposed CovSegNet architecture, several state-of-the-art networks are considered. To compare on a fair platform, most of these networks are implemented using their open-source implementation, and same train-test folds are used for performance evaluation. Infection segmentation performances using slice-based 2-D operations and volumetric 3-D operations are summarized in Tables VI and VII, respectively. CovSegNet2D provides a 4.2% higher dice score in Database-1, and an 8.6% improvement in dice score in Database-2 compared to the second-highest score (Semi-Inf-Net). Hence, consistent improvements in performances have been achieved in 2-D slice based analysis using CovSegNet2D. Moreover, in the volumetric analysis approach, CovSegNet3D provides 8.4% higher dice score and 9.4% higher IoU score compared to the next-best performing model (VNet). Thus, the 3-D variant of CovSegNet provides consistent improvements over other 3-D counterparts of existing networks. It should be noticed that the proposed hybrid scheme combining CovSegNet2D and CovSegNet3D provides the most optimum performance with a dice score of 94.1% and IoU score of 90.2%. Some of the qualitative visualizations of performances obtained in different challenging conditions are shown in Fig. 10. For having the volumetric information of the Database-1, the proposed hybrid scheme is employed here, whereas only 2-D slice based analysis is carried out in Database-2 using CovSegNet2D. It should be noted that the proposed scheme performs consistently better compared to other networks in segmenting most of the challenging diffused, blurred, and varying shaped edges of COVID lesions. Moreover, quantitative performances on challenging multiclass lesion segmentation, including separate ground-glass opacity (GGO) and consolidation regions, are summarized in Table IX, where 8.2% improvement in dice score is obtained in GGO segmentation and 11% improvement in consolidation segmentation using CovSegNet architecture over the other best-performing approaches. Additionally, from the visual analysis of the performances shown in Fig. 11, it can be easily noted that the proposed network considerably reduces the false predictions even in these challenging conditions compared to other state-of-the-art approaches.

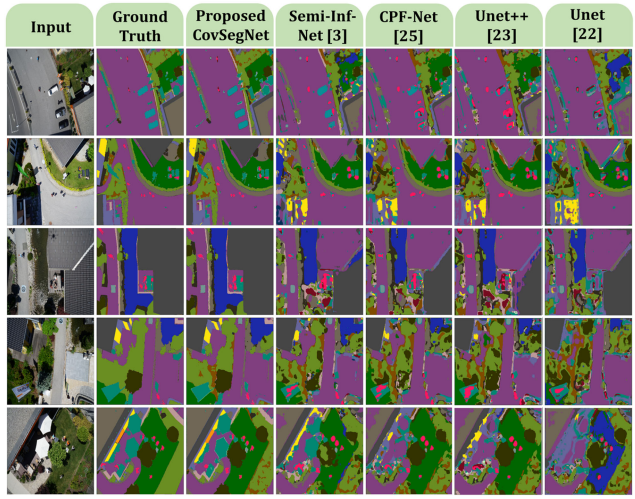

Fig. 12. Visual representations of the semantic segmentation of drone images from Database-3 obtained using different state-of-the-art networks.

Furthermore, quantitative results obtained from nonclinical Database-3 are summarized in Table X, which shows the significant performance improvement with 22.4% improvement in dice score, and 21.8% improvement in mean IoU compared to the Unet architecture. Weighted mean performances over all 20 classes are taken for better estimation. In Fig. 12, visual representations of some of the sample images are shown for different networks in Database-3, which more conspicuously signifies the better performance of the proposed architecture. Since Database-3 is very complicated with a huge number of classes, the performance differences between the proposed CovSegNet and other existing networks are more prominent as this dataset demands effective exploitation of minute, complex, and scattered features of diversified classes.

E. Computational Efficiency Analysis of Numerous Approaches

The proposed CovSegNet architecture ensures the proper optimization of all the network parameters through improved parallelization that enhances efficient gradient propagation in the whole network. However, this improved parallelism also poses some computational burden for the effective exploitation of the network parameters. Nevertheless, the CovSegNet architecture provides additional opportunity for horizontal scaling as well as vertical scaling that facilitates the performance improvement with much shallower variant. On the contrary, other traditional networks solely depend on vertical scaling that exponentially increases the computational burden with exponential increase

TABLE XI
COMPUTATIONAL EFFICIENCY ANALYSIS OF NUMEROUS ARCHITECTURES ALONG WITH THE PERFORMANCES OBTAINED ON DATASET-1

2-D analysis						3-D analysis					
Architecture	Details	Total Parameters(M)	GPU Usage(GB)	Inference Time(s)	Mean Dice(%)	Architecture	Details	Total Parameters(M)	GPU Usage(GB)	Inference Time(s)	Mean Dice(%)
Unet [22]	-	31.0	2.1	0.10	82.3	**Unet3D** [22]	-	90.3	13.2	1.22	84.2
Semi-Inf-Net [12]	-	33.3	6.8	0.18	86.9	**Vnet3D** [19]	-	45.1	15.1	1.16	85.7
Unet++ [28]	-	27.0	6.5	0.17	84.1	**MultiResUnet3D** [27]	-	18.1	12.9	1.15	84.5
CPF-Net [30]	-	32.4	2.3	0.12	84.4	**Attention Unet3D** [29]	-	103.5	20.9	1.13	85.9
CovSegNet2D-v1 (ours)	L-2, S-2	0.37	1.1	0.05	75.3	**CovSegNet3D-v1 (ours)**	L-2, S-2	1.1	7.0	1.02	79.8
CovSegNet2D-v2 (ours)	L-3, S-2	1.60	1.8	0.07	85.8	**CovSegNet3D-v2 (ours)**	L-3, S-2	4.6	13.7	1.21	89.2
CovSegNet2D-v3 (ours)	L-4, S-2	6.70	3.3	0.11	89.6	**CovSegNet3D-v3 (ours)**	L-4, S-2	19.0	22.2	1.85	92.3
CovSegNet2D-v4 (ours)	L-5, S-2	27.0	7.0	0.20	91.1	**CovSegNet Hybrid (ours)**	2D(L-4, S-2) 3D(L-2,S-2)	7.8	10.5	1.14	94.1

of the number of convolutional filters in the deeper layers. In Table XI, the computational efficiency of different networks are summarized, where performances of different variants of CovSegNet is summarized based on the number of levels (L) and stages (S). For 2-D processing, it is to be noted that the CovSegNet2D-v2 achieves 3.5% higher dice score compared to the Unet while incorporating only three-levels (L-3), and two-horizontal stages (S-2). Due to lower number of filtering operations in the upper vertical levels, significantly lower number of parameters (reduced 94.8%) are incorporated. However, for proper optimization of these parameter with improved parallelism in the network, comparatively lower gain is achieved in terms of the GPU consumption (reduced 14.2%) and inference time (reduced 30%) with respect to the Unet. A similarobservation can be carried out for 3-D analysis with CovSegNet3D. It is clear that 3-D processing increases computational complexity greatly compared to the 2-D networks. However, it should be noticed that CovSegNet-Hybrid provides the best achievable dice score (94.1%) while consisting of 0.09x parameters of Unet3D with 0.08s reduction of inference time. This significant reduction in parameter counts is mainly achieved by integrating a shallower variant of CovSegNet3D with the CovSegNet2D. Moreover, this hybrid processing effectively extracts both the interslice and intraslice contextual information that are responsible for the highest dice score. Therefore, this hybrid scheme provides considerable advantages over other existing 3-D variants in terms of parameters, and dice scores with comparable processing speed.

F. Discussions, Limitations, and Future Studies

In summary, numerous architectural renovations assist in achieving state-of-the-art performance on COVID lesion segmentation. The horizontal and vertical expansion mechanisms provide the opportunity to incorporate more detailed features as well as more generalized features, which improved the feature quality considerably that is particularly effective in distinguishing multiclass, scattered COVID lesions with widely varied shapes. Moreover, the improved gradient flow throughout the network, achieved with the introduction of MSF module and scale transition modules, have greatly reduced the contextual information loss in the generalization process and have also ensured the best optimization of all network parameters that particularly contribute to recover and distinguish the blurry, diffused edges of COVID lesions as well as the very minute instances of abnormalities. Furthermore, the integration of a hybrid 2-D–3-D networking scheme exploits both the intraslice and interslice contextual information without increasing computational burden that results in more precise, finer segmentation performance mostly in challenging conditions.

Although consistent performances have been achieved in both the datasets for COVID lesion segmentation, this study should be carried on larger datasets consisting of wide variations of subjects. However, in the current conditions of the pandemic, it is difficult to gather a considerably higher amount of data. The study proposed in this article will be extended with the incorporation of diversified datasets, including patient-based study considering age, sex, health conditions, and geographical locations of the patients. Due to the novel characteristics of the COVID infections, it is difficult to predict the risk and vulnerability among diverse subjects that can be effective for reducing the spread and better prevention. An in-depth, closer, patient-specific study should be carried out for better understandings of the nature of the infection. Moreover, generative adversarial network-based optimization can be carried out to generate more amount of realistic, synthetic data to overcome the limitations of available data. Additionally, this scheme is supposed to be extended for incorporating automated segmentation-classification joint optimization along with the severity prediction scheme of COVID infections.

IV. CONCLUSION

In this article, an automated scheme was proposed with an efficient neural network architecture (CovSegNet) for very precise lung lesion segmentation of COVID CT scans that provides outstanding performances with 8.4% average improvement of dice score over two datasets. The introduced scale transition operations were found to be very effective for replenishing contextual information loss through repeated integration of generated multiscale features in both upscaling and downscaling operations. It was found that horizontal expansion mechanism with multistage encoder–decoder modules assists in further improvements for gathering more multiscale contextual information when coupled with the traditional vertical expansion mechanism. Moreover, the MSF module with a PF scheme not only substantially reduced the semantic gaps between subsequent encoder–decoder modules but also introduced parallel interlinking among multiscale features that greatly mitigates the vanishing gradient issues for better optimization. Furthermore, the two-phase optimization scheme with hybrid 2-D–3-D processing provides considerable improvement over traditional single domain approaches for introducing more contextual information to gather finer details. It was shown that the proposed scheme is capable of segmenting

infected regions along with multiclass COVID-19 lesions with unprecedented precision even in challenging conditions with blurred, diffused, and scattered edges. Moreover, it was found that the proposed network is not only effective in COVID lesion segmentation but also provides state-of-the-art performance on a nonclinical, challenging, multiclass semantic segmentation task that proves the wide applicability of the proposed scheme. Therefore, the proposed scheme can be easily optimized on numerous applications that can be an effective alternative to other state-of-the-art approaches.

REFERENCES

[1] V. Surveillances, "The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19)-China, 2020," *China CDC Weekly*, vol. 2, no. 8, pp. 113–122, 2020.

[2] C. Wang, P. W. Horby, F. G. Hayden, and G. F. Gao, "A novel coronavirus outbreak of global health concern," *Lancet*, vol. 395, no. 10223, pp. 470–473, 2020.

[3] W. J. Guan *et al.*, "Clinical characteristics of coronavirus disease 2019 in China," *New England J. Med.*, vol. 382, no. 18, pp. 1708–1720, 2020.

[4] T. Ai *et al.*, "Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: A report of 1014 cases," *Radiology*, vol. 296, no. 2, pp. E32–E40, 2020.

[5] H. Kim, H. Hong, and S. H. Yoon, "Diagnostic performance of CT and reverse transcriptase polymerase chain reaction for coronavirus disease 2019: A meta-analysis," *Radiology*, vol. 296, no. 3, pp. E145–E155, 2020.

[6] L. Li *et al.* "Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT," *Radiology*, vol. 296, 2020, Art. no. 200905.

[7] R. Vaishya, M. Javaid, I. H. Khan, and A. Haleem, "Artificial intelligence (AI) applications for COVID-19 pandemic," *Diabetes Metabolic Syndrome: Clinical Res. Rev.*, vol. 14, no. 4, pp. 337–339, 2020.

[8] P. Huang *et al.*, "Use of chest CT in combination with negative RT-PCR assay for the 2019 novel coronavirus but high clinical suspicion," *Radiology*, vol. 295, no. 1, pp. 22–23, 2020.

[9] T. Mahmud, M. A. Rahman, and S. A. Fattah, "CovXNet: A multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest X-ray images with transferable multi-receptive feature optimization," *Comput. Biol. Med.*, vol. 122, 2020, Art. no. 103869.

[10] Y. Oh, S. Park, and J. C. Ye, "Deep learning COVID-19 features on CXR using limited training data sets," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2688–2700, Aug. 2020.

[11] H. Kang *et al.* "Diagnosis of coronavirus disease 2019 (COVID-19) with structured latent multi-view representation learning," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2606–2614, Aug. 2020.

[12] D.-P. Fan *et al.*, "Inf-Net: Automatic COVID-19 lung infection segmentation from CT images," *IEEE Trans. Med. Imag.*, vol. 39, no. 8, pp. 2626–2637, Aug. 2020.

[13] N. Saeedizadeh, S. Minaee, R. Kafieh, S. Yazdani, and M. Sonka, "COVID TV-UNet: Segmenting COVID-19 chest CT images using connectivity imposed U-net," 2020, *arXiv:2007.12303*.

[14] T. Zhou, S. Canu, and S. Ruan, "An automatic COVID-19 CT segmentation network using spatial and channel attention mechanism," 2020, *arXiv:2004.06673*.

[15] Q. Yan *et al.*, "COVID-19 chest CT image segmentation—A deep convolutional neural network solution," 2020, *arXiv:2004.10987*.

[16] Y. Qiu, Y. Liu, and J. Xu, "MiniSeg: An extremely minimum network for efficient COVID-19 segmentation," 2020, *arXiv:2004.09750*.

[17] J. Ma *et al.*, "Towards efficient COVID-19 CT annotation: A benchmark for lung and infection segmentation," 2020, *arXiv:2004.12537*.

[18] D. Müller, I. S. Rey, and F. Kramer, "Automated chest CT image segmentation of COVID-19 lung infection based on 3D U-net," 2020, *arXiv:2007.04774*.

[19] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis.*, 2016, pp. 565–571.

[20] L. Zhang *et al.*, "Block level skip connections across cascaded V-net for multi-organ segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 9, pp. 2782–2793, Sep. 2020.

[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.

[23] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data," *ISPRS J. Photogrammetry Remote Sens.*, vol. 162, pp. 94–114, 2020.

[24] M. Al Ghamdi, M. Abdel-Mottaleb, and F. Collado-Mesa, "DU-Net: Convolutional network for the detection of arterial calcifications in mammograms," *IEEE Trans. Med. Imag.*, vol. 39, no. 10, pp. 3240–3249, Oct. 2020.

[25] Z. Zeng, W. Xie, Y. Zhang, and Y. Lu, "RIC-Unet: An improved neural network based on Unet for nuclei segmentation in histology images," *IEEE Access*, vol. 7, pp. 21420–21428, 2019.

[26] L. Mou, L. Chen, J. Cheng, Z. Gu, Y. Zhao, and J. Liu, "Dense dilated network with probability regularized walk for vessel detection," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1392–1403, May 2020.

[27] N. Ibtehaz and M. S. Rahman, "MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Netw.*, vol. 121, pp. 74–87, 2020.

[28] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.

[29] O. Oktay *et al.* "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[30] S. Feng *et al.*, "CPFNet: Context pyramid fusion network for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 10, pp. 3008–3018, Oct. 2020.

[31] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3D fully convolutional deep networks," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, 2017, pp. 379–387.

[32] N. Abraham and N. M. Khan, "A novel focal Tversky loss function with improved attention U-Net for lesion segmentation," in *Proc. IEEE 16th Int. Symp. Biomed. Imag.*, 2019, pp. 683–687.

[33] "*COVID-19 CT lung and infection segmentation dataset*," 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3757476

[34] "*COVID-19 CT segmentation dataset*," 2020. [Online]. Available: https://medicalsegmentation.com/covid19/

[35] "*Semantic drone dataset*," 2019. [Online]. Available: https://www.tugraz.at/index.php?id=22387/