Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
COVID TV-Unet: Segmenting COVID-19 Chest CT Images Using Connectivity Imposed Unet

Narges Saeedizadeh, Shervin Minaee, Rahele Kafieh, Shakib Yazdani, Milan Sonka

PII: S2666-9900(21)00006-9
DOI: https://doi.org/10.1016/j.cmpbup.2021.100007
Reference: CMPBUP 100007

To appear in: Computer Methods and Programs in Biomedicine Update

Received date: 8 December 2020
Revised date: 13 March 2021
Accepted date: 29 March 2021

Please cite this article as: Narges Saeedizadeh, Shervin Minaee, Rahele Kafieh, Shakib Yazdani, Milan Sonka, COVID TV-Unet: Segmenting COVID-19 Chest CT Images Using Connectivity Imposed Unet, Computer Methods and Programs in Biomedicine Update (2021), doi: https://doi.org/10.1016/j.cmpbup.2021.100007

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Highlights

- A segmentation framework to detect infected chest regions in CT images
- A regularization term based on 2D-anisotropic total-variation is added to the loss function
- A relatively largescale CT segmentation dataset of around 900 images
- Identifying infected regions with mIoU rate of 99%, and a Dice score of 86%
COVID TV-Unet: Segmenting COVID-19 Chest CT Images Using Connectivity Imposed Unet

Narges Saeedizadeh1,∗, Shervin Minaei2,∗, Rahele Kafeh1,∗, Shabih Yazdani3, Milan Sonka4
1Medical Image and Signal Processing Research Center, Isfahan University of Medical Sciences, Iran
2Snap Inc., Seattle, WA, USA
3ECE Department, Isfahan University of Technology, Iran
4Iowa Institute for Biomedical Imaging, The University of Iowa, Iowa City, USA
∗Corresponding Authors

Abstract—The novel corona-virus disease (COVID-19) pandemic has caused a major outbreak in more than 200 countries around the world, leading to a severe impact on the health and life of many people globally. By October 2020, more than 44 million people were infected, and more than 1,000,000 deaths were reported. Computed Tomography (CT) images can be used as an alternative to the time-consuming RT-PCR test, to detect COVID-19. In this work we propose a segmentation framework to detect chest regions in CT images, which are infected by COVID-19. An architecture similar to a Unet model was employed to detect ground glass regions on a voxel level. As the infected regions tend to form connected components (rather than randomly distributed voxels), a suitable regularization term based on 2D-anisotropic total-variation was developed and added to the loss function. The proposed model is therefore called ”TV-Unet”. Experimental results obtained on a relatively large-scale CT segmentation dataset of around 900 images, incorporating this new regularization term leads to a 2% gain on overall segmentation performance compared to the Unet trained from scratch. Our experimental analysis, ranging from visual evaluation of the predicted segmentation results to quantitative assessment of segmentation performance (precision, recall, Dice score, and mIoU) demonstrated great ability to identify COVID-19 associated regions of the lungs, achieving a mIoU rate of over 99%, and a Dice score of around 86%.

Keywords: Deep Learning, Computed Tomography, COVID-19, Image Segmentation, Convolutional Encoder Decoder, Total Variation

I. INTRODUCTION

Since December 2019, a novel corona-virus (SARS-CoV-2) has spread from Wuhan to the whole China, and then to many other countries. At the end of January 2020, the World Health Organization (WHO) declared COVID-19 a Public Health Emergency of International Concern [1]. By October 2020, more than 44 million confirmed cases, and more than 1,000,000 deaths were reported across the world [2]. While infection rates are decreasing in some countries, numbers of new infections continue quickly growing in many other countries, signaling the continuing and global threat of COVID-19 [3]–[5].

Up to this point, no effective treatment has yet been proven for COVID-19. Therefore, for prompt prevention of COVID-19 spread, accurate and rapid testing is extremely pivotal. The reverse transcription polymerase chain reaction (RT-PCR) has been considered the gold standard in diagnosing COVID-19. However, the shortage of available tests and testing equipment in many areas of the world limits rapid and accurate screening of suspected subjects. Even under best circumstances, obtaining RT-PCR test results takes more than 6 hours and the routinely achieved sensitivity of RT-PCR is insufficient [6]. On the other hand, the radiological imaging techniques like chest X-rays and computed tomography (CT) followed by automated image analysis [7] may successfully complement RT-PCR testing. CT screening provides three-dimensional view of the lung and is therefore more sensitive (although less widely available) compared to chest X-ray radiography.

In a systematic review [8] the authors indicated that CT images are sensitive in detection of COVID-19 before observation of some clinical symptoms. Typical signs of COVID-19 in CT images consist of unilateral, multifocal and peripherally based ground glass opacities (GGO), interlobular septal thickening, thickening of the adjacent pleura, presence of pulmonary nodules, round cystic changes, bronchiectasis, pleural effusion, and lymphadenopathy [9], [10]. Accurate and rapid detection and localization of these pathological tissue changes is critical for early diagnosis and reliable severity assessment of COVID-19 infection. As the number of affected patients is high, manual annotation by well-trained expert radiologists is time consuming, subject to inter- and intra-observer variability, and slows down the CT analysis. The urgent need for automatic segmentation of typical COVID-19 CT signatures is widely appreciated and deep learning methods can offer a unique solution for identifying COVID-19 signs of infection in clinical-quality images frequently suffering from variations in CT acquisition parameters and protocols [11].

In this work, we present a deep learning based framework for automatic segmentation of pathologic COVID-19-associated tissue areas from clinical CT images available from publicly available COVID-19 segmentation datasets. Our solution is based on adapting and enhancing a popular deep learning medical image segmentation architecture Unet for COVID-19 segmentation task. As COVID-19 tissue regions tend to form connected regions identifiable in individual CT slices, a "connectivity promoting regularization" term was added to the specifically designed training loss function to encourage the model to prefer sufficiently large connected segmentation regions of desirable properties.

The main contributions of our work can be summarized as follows:

I. Archi...
• Development of a novel connectivity-promoting regularization loss function for an image segmentation framework detecting pathologic COVID-19 regions in pulmonary CT images,

• Quantitative validation showing improved performance attributable to our new TV-Unet approach compared to published state-of-the-art segmentation approaches

• Public sharing the developed software code facilitating research and medical community use (https://github.com/narges-sa/COVID-CT-Segmentation)

This paper is therefore a novel contribution in the rapidly advancing effort to develop techniques and approaches for reliable detection and quantification of COVID-19 disease in human lungs affected by the disease. Our work brings additional advances attributable to the new TV-Unet approach and demonstrates the performance improvement of our approach compared to recently reported approaches.

II. RELATED WORKS

COVID-19 segmentation from CT images has recently received focused attention of the medical imaging community. There has been a huge progress in the performance of image segmentation models using various deep learning frameworks in recent years [12]–[17]. In [18], Fan et al. proposed InfNet, to identify infected regions from chest CT slices. In their proposed model, a parallel partial decoder is used to aggregate high-level features and generate a global map. Then, the implicit reverse attention and explicit edge-attention are utilized to model the boundaries and enhance the representations. Unfortunately, this model used a very small dataset of CT-labeled images for segmentation, which consisted of only 100 CT slices, making it hard to generalize and compare their result. In order to have a fair comparison, results from TV-Unet and other methods which were implemented in the InfNet, namely Semi-InfNet, and Semi-InfNet+FCN8s [18], Unet++ [19], DeepLab-v3 [20], and FCN8s [21] are provided in the Discussion. In [22], Elharrouss et al. proposed an encoder-decoder based model for lung infection segmentation using CT-scan images. The proposed model initially considers the image structure and texture to extract the ROI of infected area, and then uses the ROI along with the image structure to predict the infected regions. Elharrouss et al. also trained this model on a small dataset of CT images, and achieved reasonable performance. In [23], Ma et al. prepared a new benchmark of 3D CT data with 20 cases that contained 1800+ annotated slices and provided several pre-trained baseline models that facilitated for out-of-the-box 3D segmentation. In [24], a novel COVID-19 diagnosis system was reported, utilizing a joint approach to deep image classification and segmentation (JCS). In their work, a large-scale COVID-19 dataset is used for classification and segmentation purposes (COVID-CS dataset) which contains 144,167 chest CT images of 400 COVID-19 subjects and 350 uninfected cases. They reached Dice score of 78.5% on the segmentation test set.

III. THE PROPOSED FRAMEWORK

Despite a large number of patients suffering from COVID-19, despite a growing number of COVID-19 volumetric CT scans, the availability of labeled CT images that can be used for training of deep learning methods is still limited. Therefore, our strategy relies heavily on the use of transfer learning, initiating the training from a model previously developed for medical image segmentation (segmentation of neuronal structures in electron microscopic stacks), and adapt it toward this task. To better suit the segmentation task at hand, we employed an architecture similar to the Unet, one of the most successful deep learning medical image segmentation approaches, and modified its loss function to prefer COVID-19 specific foreground mask connectivity.

A. Unet Architecture

Unet is one of the popular segmentation models which is based on encoder-decoder neural architecture and use of skip connections, and was originally proposed by Ronneberger et al. [17]. The network architecture of Unet is illustrated in Fig. 1. In the encoder part, model gets an image as input and applies multiple layers of convolution, max-pooling and ReLU activation and compresses the data into a latent space. In the decoder part, the network attempts to decode the information from the latent space using transposed convolution operation (deconvolution) and produce the segmentation mask of the image. The rest of the operations are similar to the aforementioned ones in the encoder part. One difference between Unet and plain encoder-decoder model is the use of skip-connections to send the information from the corresponding high-resolution layers of the encoder to the decoder, which can help the network to better capture small details that are present in high-resolution. Figure 1 illustrates the general architecture of a Unet model.

Fig. 1. The architecture of Unet model

Similar to other neural segmentation models, Unet uses the loss function in Eq 1 during training.

\[ L_{U\text{net}} = \sum_{x \in \Omega} \omega(x) \log(p(x|z(x))) \]  

(1)

Where \( \ell : \Omega \to \{1, \ldots, K\} \) considering that \( \Omega \subset \mathbb{Z}^2 \) and \( K \) denotes the total number of classes. When using the binary
cross-entropy, K=2. Moreover, the soft-max is defined as
\[ p_k(x) = \frac{\exp(a_k(x))}{\sum_{k=1}^{K} \exp(a_k(x))} \] where \( a_k(x) \) represents the activation in feature channel K [17].

Additionally, in a network that contains many convolutional layers, the initial weights greatly affect the performance. In this regard, approximately unit-variances of the initial weights of network feature maps performed best. In the network with convolutional and ReLU layers such as Unet, using a Gaussian distribution with a standard deviation of \( \sqrt{2/N} \) is an effective solution. In this method, \( N \) represents the number of incoming nodes of one neuron [25].

### B. Connectivity Regularization

The segmentation maps usually consist of a number of connected components, and single-pixel regions are rare. To encourage our segmentation model to generate segmentation maps with connected components of desirable sizes, we found that incorporating an explicit regularization term in the training loss function can greatly improve connectivity requirements for the predicted segmentation regions. It is worth noting that Unet trained from scratch can also implicitly learn such behavior from training data to some extent, assuming sufficient data sizes are available, which is not quite the case in our situation. Several strategies were developed and considered to impose desired connectivity requirements within images such as adding group-sparsity or incorporate total variation terms [26]–[28]. Based on achieved experience, we decided to use total variation (TV) as its gradient update is computationally attractive during the backward pass stage.

TV penalizes the generated images with large variations among neighboring pixels, leading to more connected and smoother solutions [27]. Total variation of 1D discrete signals (\( y = [y_1, ..., y_N] \)) is straightforward, and can be defined as:

\[ TV(y) = \sum_{n=1}^{N} |y_{n+1} - y_n|, \] (3)

where \( D \) is a \((N-1) \times N\) matrix as below:

\[
D = \begin{bmatrix}
1 & -1 & 0 & \ldots & 0 \\
0 & 1 & -1 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & 1 & -1
\end{bmatrix}.
\]

For 2D signals (\( Y = [y_{i,j}] \)), we can use isotropic or anisotropic versions of 2D total variation [26]. To simplify our optimization problem, we have used the anisotropic version of TV, which is defined as the sum of horizontal and vertical gradients at each pixel:

\[ TV(Y) = \sum_{i,j} |y_{i+1,j} - y_{i,j}| + |y_{i,j+1} - y_{i,j}|. \] (4)

In our case we can add the total variation of the predicted binary mask for COVID-19 pixels to the loss function. Adding this 2D-TV regularization term to our framework will promote the connectivity of the produced segmentation regions. The new loss function for our model would then be defined as:

\[ L_{TV-Unet} = L_{Unet} + \lambda TV(M(x)), \] (5)

where \( L_{Unet} \) is the binary cross-entropy loss, \( M(x) \) is a COVID-19 probability mask generated in the last layer of the network, and \( \lambda \) is a constant that needs to be tuned. In this paper, based on the model performance on the validation and test set, \( \lambda \) was set as \( 1/(255^{\text{number of pixels}}) \) to provide the model with balance weights.

### IV. COVID CT Segmentation Dataset

We have combined all available data from the COVID-19 CT segmentation dataset [29], consisting of 929 CT slices from 49 patients. Out of these, 473 CT-image slices are labeled as including COVID-19 pathologies with Ground-Glass pathology regions identified by expert tracing. The remaining 456 CT image slices are labeled as COVID-19 pathology-free. CT-slice sizes were either 512×512 or 630×630. While a small subset of the 929 CT images also have regions of additional pathologies identified and labeled as Consolidation and/or Pleural Effusion, these pathologies were not consistently available for all the data. After consulting with a board-certified radiologist (GS) who confirmed that Ground-Glass pathology is most relevant for detecting COVID-19, this work focuses on the Ground Glass mask, and does not consider the Consolidation and Pleural Effusion masks due to their small numbers and lack of consistency across the dataset.

Some sample images from the used dataset are shown in Figure 2 and 3. Figure 2 demonstrates the difference between normal and COVID-19 images. The images in Figure 3 denote the original images, and their corresponding COVID-19 Ground-Glass masks. The red boundary contours in the second row are drawn to better show the parts containing COVID-19, and are not a part of the original image.

**Fig. 2.** The difference between normal and COVID-19 images

**A. Training, Validation, and Testing Sets**

To evaluate the effect of radically different training/testing set composition and to demonstrate the robustness of the obtained results, two different splits of training, validation, and testing sets are selected from this dataset (Table I).
Fig. 3. Sample images from the COVID-19 CT segmentation dataset. The first row shows two COVID-19 images. The red boundary contours in the second row denote regions of COVID-19 Ground-Glass pathology and are not a part of the original image data. The third row shows Ground-Glass masks.

In Split 1, data from a relatively large number of 46 training/validation-set patients and a small number of only 3 testing-set patients were used. 729 CT image slices formed the training and validation sets, and 200 images the testing set. In Split 2, a more balanced distribution of patient numbers was used with 654 CT image slices from 35 patients included in the training and validation sets, and 275 images from 14 patients formed the testing set.

| Data     | Number of Images in Split 1 | Number of Images in Split 2 |
|----------|----------------------------|-----------------------------|
| Training | 654                        | 590                         |
| Validation | 75                        | 64                          |
| Test     | 200                        | 275                         |
| Total    | 929                        | 929                         |

V. EXPERIMENTAL RESULTS

In this section we provide a detailed experimental analysis of the proposed segmentation framework, by presenting both qualitative and quantitative results as well as comparing our results with a baseline approach.

A. Evaluation Metrics

There are several metrics that are used by the research community to measure the performance of segmentation models, including precision, recall, Dice coefficient and mean Intersection over Union (mIoU). These metrics are also widely used in medical domain, and are defined as below.

**Precision** is calculated as the ratio of pixels correctly predicted as COVID divided by total pixels predicted as COVID-19

\[
\text{Precision} = \frac{TP}{TP + FP},
\]

where TP refers to the true positive (the number of correctly labeled COVID-19 pixels) and FP refers to the false positive (the number of incorrectly labeled COVID-19 pixels).

**Recall** is the ratio of pixels correctly labeled as COVID-19 divided by total number of actual COVID-19 pixels and is defined as

\[
\text{Recall} = \frac{TP}{TP + FN},
\]

where TP is the number of false positives, and FN refers to the false negatives and is the number of pixels mistakenly labeled as non-COVID.

Precision-Recall (PR) curve is a popular way to holistically look at the model performance as a plot of the precision (y-axis) versus the recall (x-axis) rates for different thresholds.

**Dice Coefficient** (also known as Dice score or DSC) is another popular metric especially for the multi-class image segmentation:

\[
\text{DSC} = \frac{2|A \cap B|}{|A| + |B|},
\]

where A and B denote the predicted and ground-truth masks.

**Intersection over Union** (also known as the Jaccard index) is used to evaluate the similarity between ground truth and predicted segmentation masks. It is defined as the size of the intersection divided by the size of the union of the target mask and predicted segmentation map

\[
\text{IoU}(A, B) = \frac{|A \cap B|}{|A \cup B|},
\]

where A and B are the predicted and ground-truth masks. **Mean-IoU** is the average IoU value over all classes. If A and B are both empty, IoU(A,B) is defined as 1. IoU ranges between 0 and 1.

It is worth mentioning that Dice coefficient and IoU are positively correlated.

B. Model Hyper-parameters

Model hyper-parameters govern the machine learning processes and as such are crucial for achieving good performance, especially in the case of deep neural networks. Hyper-parameter tuning can be done in two different ways, automatically and manually. In this work, we manually evaluated different combinations of hyper-parameters and selected the best combination. To simplify the tuning process, we fixed the number of epochs to 100, and the batch-size to 32. We
designed and compared different loss functions (such as binary cross-entropy (BCE) [30], Dice coefficient loss [31], and BCE plus total variation regularization), different optimizers (such as ADAM [32], Adagrad [33], Adadelta [34] and stochastic gradient descent (SGD) [35]), and different learning rates On First Split. We used adaptive learning rate scheduling and early stopping [36] criteria as below, which achieved good performance on the validation set:

- Learning rate is decayed whenever the validation loss does not improve for 5 continuous epochs.
- Early stopping is applied whenever the validation loss does not improve for 10 continuous epochs.

Table II shows the impact of the loss function design on First Split on the model performance with binary cross-entropy and the proposed connectivity regularized loss function achieving the best performance.

### Table II. Overall Performance with Different Loss Functions Employed, the Best Cut-off Threshold of 0.3 used. Best Performance Shown in Bold Font.

| Loss           | Optimizer | Learning Rate | mIOU | DSC   | Average Precision |
|----------------|-----------|----------------|------|-------|-------------------|
| BCE            |           |                | 0.993| 0.839 |                   |
| DSC            |           | 0.990          | 0.764| 0.90  |                   |
| BCE+DSC        | ADAM      | 0.993          | 0.843| 0.91  |                   |
| BCE+TV         | ADAM      | 0.995          | 0.864| 0.94  |                   |

The impact of the optimizer on the model performance is shown in Table III. As we can see, ADAM achieves the highest performance in terms of all metrics.

### Table III. Overall Performance for Different Optimizer Selection, using the Best Cut-off Threshold of 0.3. Best Performance Shown in Bold Font.

| Loss | Optimizer | Learning Rate | mIOU | DSC   | Average Precision |
|------|-----------|----------------|------|-------|-------------------|
| BCE+TV| ADAM      | 0.001          | 0.995| 0.684 | 0.86              |
|      | SGD       |                | 0.985| 0.573 | 0.8               |
|      | Adadelta  | 0.991          | 0.780| 0.9    |                   |
|      | Adagrad   | 0.992          | 0.784| 0.9    |                   |

Table IV provides the analysis of model performance for two different learning rate values when using (the best performing) ADAM optimization.

### Table IV. Overall Model Performance for Different Learning Rates, again for the Best Cut-off Threshold of 0.3. Best Performance Shown in Bold Font.

| Loss       | Optimizer | Learning Rate | mIOU | DSC   | Average Precision |
|------------|-----------|----------------|------|-------|-------------------|
| BCE+TV     | ADAM      | 0.0001         | 0.993| 0.838 | 0.92              |

Table V demonstrates the impact of varying the value of $\lambda$ on the final result when using (the best performing) ADAM optimization with the learning rate of 0.001.

Note that the model predicts a probability for each pixel, showing its likelihood of belonging to the pathologic COVID-19 region (zero denotes Non-COVID pixels and one denotes COVID-19 pathology). These probabilities are thresholded, different thresholds yield certain sensitivity/specificity rates. Threshold value of 0.3 achieved the best performance on the validation set and was therefore used to report the results of the proposed model. The impact of modifying the threshold values on the model accuracy is given in Section V-D.

### C. Predicted Masks

Qualitative result showing how close our predicted masks are to the ground-truth masks are given in Figure 4 for 5 sample images from the test set. As can be seen when the desired region is very tiny, the Unet trained from scratch cannot distinguish the segmentation region and background very well, while the proposed TV-Unet model performs notably better. If, however, the GT-mask consists of only a few isolated points (e.g., the second row in Figure 4), our model sometimes fails detecting such points. Clearly, such isolated points do not provide a strong-enough supporting information for local detection of COVID-19 infection.

![Fig. 4. Predicted segmentation masks by Unet trained from scratch and the proposed TV-Unet for a typical sample images from the testing set.](image-url)
D. Cut-off Threshold Impact on Model Performance

As discussed previously, our model predicts a probability score for each pixel, showing the likelihood of its being in a COVID-19 pathology region. Different cut-off thresholds can be used on those probabilities to decide COVID-19 labeling. By increasing the cut-off threshold, less and less pixels would be labeled as COVID-19 pathology. Tables VI and VII show the model performance (in terms of precision, recall, and mIoU) for eight different values of cut-off thresholds for the Split-1 and Split-2 datasets. The cut-off threshold of 0.3 results in the highest Dice score, and mIoU metric, and therefore was employed to compare our model with other baseline models.

| TABLE VI. PRECISION, RECALL, DICE SCORE, AND mIoU RATES OF TV-UUNET MODEL FOR DIFFERENT THRESHOLD VALUES FOR SPLIT 1. |  |
|---|---|---|---|---|
| Threshold | Recall | Precision | mIoU | DSC |
| 0.1 | 0.98 | 0.70 | 0.99 | 0.83 |
| 0.2 | 0.91 | 0.64 | 0.98 | 0.87 |
| 0.3 | 0.87 | 0.67 | 0.99 | 0.86 |
| 0.4 | 0.83 | 0.70 | 0.99 | 0.85 |
| 0.5 | 0.74 | 0.75 | 0.99 | 0.83 |
| 0.6 | 0.66 | 0.80 | 0.98 | 0.78 |
| 0.7 | 0.54 | 0.85 | 0.97 | 0.72 |
| 0.8 | 0.42 | 0.90 | 0.96 | 0.57 |

| TABLE VII. PRECISION, RECALL, DICE SCORE, AND mIoU RATES OF TV-UUNET MODEL FOR DIFFERENT THRESHOLD VALUES FOR SPLIT 2. |  |
|---|---|---|---|---|
| Threshold | Recall | Precision | mIoU | DSC |
| 0.1 | 0.89 | 0.67 | 0.98 | 0.75 |
| 0.2 | 0.83 | 0.70 | 0.99 | 0.74 |
| 0.3 | 0.78 | 0.75 | 0.99 | 0.74 |
| 0.4 | 0.73 | 0.77 | 0.99 | 0.75 |
| 0.5 | 0.68 | 0.80 | 0.99 | 0.74 |
| 0.6 | 0.61 | 0.85 | 0.99 | 0.72 |
| 0.7 | 0.53 | 0.90 | 0.98 | 0.69 |
| 0.8 | 0.42 | 0.95 | 0.97 | 0.66 |

To see the holistic view of the proposed model performance on all possible threshold values, Figures 5 and 6 provide the precision-recall curves on the test sets in Split 1 and Split 2, respectively. Figure 5 shows average precision of 0.92 for the Unet trained from scratch and 0.84 for our TV-Unet for Split 1 dataset (an improvement of around 0.08 in terms of average-precision). Figure 6 shows average precision of 0.67 for the Unet trained from scratch and 0.88 for our TV-Unet for Split 2, a relative improvement of 31%.

E. Model Performance Comparison with Unet trained from scratch

For a fair comparison between the proposed TV-Unet model and the Unet trained from scratch, corresponding cut-off thresholds were identified for similar recall rates for each model and compared in terms of other performance metrics. Tables VIII and IX provide the comparison between these two models for four different recall rates. Consistency, with which our TV-Unet model outperforms that of the Unet trained from scratch when considering all metrics, shows the added value of the connectivity-promoting regularization. We have an average improvement of around 2% in terms of Dice score in Split 1 and 10.9% in Split 2 studies.

F. Training Convergence Analysis

To see the model convergence during training, we provide the loss function, recall, and precision rates of the model on different epochs, in Figures 7, 8 and 9. It is worth to mention that for precision and recall, the default threshold value of 0.3.
TABLE VIII. COMPARISON OF UNET TRAINED FROM SCRATCH AND THE PROPOSED TV-UNET MODEL PERFORMANCE IN TERMS OF PRECISION, mIOU AND DSC FOR SPLIT 1. BEST PERFORMANCE FOR EACH METHOD SHOWN IN BOLD FONT.

| Model   | Recall | Precision | mIOU  | DSC  |
|---------|--------|-----------|-------|------|
| Unet    | 0.975  | 0.575     | 0.985 | 0.727|
|         | 0.945  | 0.688     | 0.990 | 0.798|
|         | 0.91   | 0.765     | 0.992 | 0.832|
|         | 0.85   | 0.834     | 0.993 | 0.841|
| TV-Unet | 0.975  | 0.675     | 0.990 | 0.798|
|         | 0.945  | 0.760     | 0.993 | 0.842|
|         | 0.91   | 0.812     | 0.994 | 0.860|
|         | 0.85   | 0.871     | 0.995 | 0.864|

TABLE IX. COMPARISON OF THE UNET TRAINED FROM SCRATCH AND TV-UNET MODEL PERFORMANCE IN TERMS OF PRECISION, mIOU AND DSC FOR SPLIT 2. BEST PERFORMANCE FOR EACH METHOD SHOWN IN BOLD FONT.

| Model   | Recall | Precision | mIOU  | DSC  |
|---------|--------|-----------|-------|------|
| Unet    | 0.810  | 0.594     | 0.983 | 0.655|
|         | 0.643  | 0.621     | 0.985 | 0.633|
|         | 0.535  | 0.655     | 0.985 | 0.595|
|         | 0.422  | 0.693     | 0.984 | 0.527|
| TV-Unet | 0.810  | 0.727     | 0.990 | 0.764|
|         | 0.643  | 0.842     | 0.990 | 0.729|
|         | 0.535  | 0.890     | 0.989 | 0.670|
|         | 0.422  | 0.926     | 0.987 | 0.580|

is used in these figures.

VI. DISCUSSION

The general lack of available pulmonary CT datasets and especially a lack of expert-annotated CT data can be seen in so-far available publications for COVID-19 pathology detection in CT images. This lack of data also forces researchers to develop 2D CT-slice-based methods rather than volumetric 3D data analyses. The entire medical imaging community is anxiously awaiting the arrival of large annotated 3D CT datasets – the analysis of which will however greatly benefit from the experience gained on current 2D CT data.

One of important accomplishments to date is a number of new approaches and publicly available – albeit small – COVID-19 CT datasets that can be used for method comparisons. One such dataset with semi-supervised COVID-19 segmentations (COVID-SemiSeg) was recently reported in [37]. The COVID-SemiSeg dataset consists of two sets. The first one contains 1600 pseudo labels generated by Semi-Inf-Net model and 50 labels by expert physicians. The second set includes 50 multi-class labels. Overall, there are 48 images that can be used for performance-comparison assessment and these CT data were used to compare our TV-Unet approach with other methods.

Our TV-Unet model reported above was therefore compared with the Inf-Net [18] and with other promising image segmentation models trained on COVID-SemiSeg dataset,
including Unet++, Semi-Inf-Net, DeepLab-v3, FCN8s, and Semi-Inf-Net+FCN8s.

Consistently using the SemiSeg dataset, Tables X, XI, and XII provide performance comparisons of individual methods in terms of recall and Dice coefficient. As can be seen from these tables, the proposed TV-Unet model outperformed all other models in all three tested experiments, which focused on a) identifying COVID-19-specific pathologic regions, Ground-Glass regions, and COVID-19 Consolidation regions.

TABLE X. COMPARISON OF THE TV-UNET MODEL PERFORMANCE WITH OTHER RECENT METHODS IN TERMS OF SENSITIVITY, SPECIFICITY AND DSC FOR PATHOLOGIC REGIONS ON COVID-SEMISEG DATASET. BEST PERFORMANCE IS SHOWN IN BOLD FONT.

| Model                | Sensitivity | Specificity | Dice Score |
|----------------------|-------------|-------------|------------|
| Unet++               | 0.672       | 0.902       | 0.518      |
| Inf-Net              | 0.692       | 0.943       | 0.682      |
| Semi-Inf-Net         | 0.725       | 0.960       | 0.759      |
| TV-Unet              | **0.808**   | **0.980**   | **0.801**  |

TABLE XI. COMPARISON OF THE TV-UNET MODEL PERFORMANCE WITH OTHER METHODS IN TERMS OF SENSITIVITY, SPECIFICITY AND DSC FOR THE GROUND-Glass MASK ON COVID-SemiSeg DATASET. BEST PERFORMANCE SHOWN IN BOLD FONT.

| Model                | Sensitivity | Specificity | Dice Score |
|----------------------|-------------|-------------|------------|
| DeepLab-v3+ (stride=8) | 0.773       | 0.823       | 0.443      |
| DeepLab-v3+ (stride=16) | **0.813**   | **0.843**   | **0.484**  |
| FCN8s                | 0.537       | 0.905       | 0.471      |
| Semi-Inf-Net+FCN8s   | 0.730       | 0.941       | 0.646      |
| TV-Unet              | **0.762**   | **0.979**   | **0.655**  |

TABLE XII. COMPARISON OF THE TV-UNET MODEL PERFORMANCE WITH SEVERAL OTHER METHODS IN TERMS OF SENSITIVITY, SPECIFICITY AND DSC FOR THE CONSOLIDATION MASK ON COVID-SemiSeg DATASET. BEST PERFORMANCE SHOWN IN BOLD FONT.

| Model                | Sensitivity | Specificity | Dice Score |
|----------------------|-------------|-------------|------------|
| DeepLab-v3+ (stride=8) | 0.100       | 0.584       | 0.117      |
| DeepLab-v3+ (stride=16) | 0.245       | 0.500       | 0.188      |
| FCN8s                | 0.212       | 0.500       | 0.221      |
| Semi-Inf-Net+FCN8s   | 0.196       | 0.670       | 0.185      |
| TV-Unet              | **0.558**   | **0.985**   | **0.537**  |

VII. CONCLUSION

A novel deep learning framework for COVID-19 segmentation from CT images was developed. We used the popular Unet architecture as the main framework, and improved its performance by an added connectivity promoting regularization term to encourage the model to generate larger contiguous segmentation maps. We showed that our trained model achieved a high accuracy rate for detecting of pathologic COVID-19 regions. We report the model performance under various hyper-parameter settings, which can be helpful for future research by the community to know the impact of different parameters on the final results. Last but not least, we demonstrated that our TV-Unet approach outperformed other reported methods. We will further extend this work to semi-supervised setting, in which a combination of labeled and unlabeled data will be used for training the model. Such an approach will undoubtedly be extremely useful as collecting accurate segmentation labels for COVID-19 remains very challenging.

ACKNOWLEDGMENT

The authors would like to thank radiologist, Dr. Ghazaleh Soufi, for her advice on the important signals in chest CT images for detecting COVID-19. We would also like to thank the providers of the publicly available pulmonary CT datasets. M. Sonka’s research effort supported, in part, by NIH grant R01-EB004640.

REFERENCES

[1] W. H. O. (2020), “Coronavirus disease (COVID-19) pandemic.” https: //www.who.int/ emergencies/diseases/novel-coronavirus-2019.

[2] https://www.worldometers.info/coronavirus/

[3] Paules, Catharine I., Hilary D. Marston, and Anthony S. Fauci. “Coronavirus infections—more than just the common cold.” Jama 323.8 (2020): 707-708.

[4] Remuzzi, Andrea, and Giuseppe Remuzzi. “COVID-19 and Italy: what next?.” The Lancet (2020).

[5] Kafieh, Rahelg, Roya Arian, Nargis Saeedizadeh, Shervin Minae, Sunil Kamar Yadav, Atefeh Vaezi, Nima Rezaei, and Shahghayegh Hashgooyo. "COVID-19 in Iran: A Deeper Look Into The Future." medRxiv (2020).

[6] Y. Fang, H. Zhang, J. Xie, M. Lin, L. Ying, P. Pang, and W. Ji, “Sensitivity of chest CT for COVID-19: Comparison to RT-PCR.” Radiology, vol. 0, no. 0, p. 200432, 0, pMID: 32073535.[Online]. Available: https://doi.org/10.1148/radiol.2020200432

[7] Minae, Shervin, Rahеле Kafieh, Milan Sonka, Shabib Yazdani, and Ghazaleh Jamalipour Soufi. "Deep-covid: Predicting Covid-19 from chest x-ray images using deep transfer learning." Medical Image Analysis, pre-pub ahead of print, doi.org/10.1016/j.medien.2020.101794 (2020).

[8] S. Salehi, A. Abedi, S. Balakrishnan, and A. Gholamrezaehad, “Coronavirus disease 2019 (COVID-19): a systematic review of imaging findings in 919 patients,” American Journal of Roentgenology, pp. 1–7, 2020.

[9] X. Ding, J. Xu, J. Zhou, and Q. Long, “Chest CT findings of COVID-19 pneumonia by duration of symptoms,” European Journal of Radiology, vol. 127, Jun. 2020, doi: 10.1016/j.ejrad.2020.109009.

[10] H. Meng et al., “CT imaging and clinical course of asymptomatic cases with COVID-19 pneumonia at admission in Wuhan, China,” J Infect, Apr. 2020, doi: 10.1016/j.jinf.2020.04.004.

[11] F. Shan, Y. Gao, J. Wang, W. Shi, N. Shi, M. Han, Z. Xue, D. Shen, and Y. Shi, “Lung infection quantification of COVID-19 in ct images with deep learning,” arXiv preprint arXiv:2003.04655, 2020.

[12] Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." IEEE transactions on pattern analysis and machine intelligence 39.12 (2017): 2481-2495.

[13] Chen, Liang-Chieh, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. "Encoder-decoder with atrous separable convolution for semantic image segmentation." In Proceedings of the European conference on computer vision (ECCV), pp. 801-818. 2018.

[14] Huang, Zilong, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. "Ccnet: Criss-cross attention for semantic segmentation." In Proceedings of the IEEE International Conference on Computer Vision, pp. 603-612. 2019.

[15] Minae, Shervin, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. "Image segmentation using deep learning: A survey." arXiv preprint arXiv:2001.05566 (2020).
[16] Minaee, Shervin, Amirali Abdolrashidi, Hang Su, Mohammed Bennamoun, and David Zhang. "Biometric recognition using deep learning: A survey." arXiv preprint arXiv:1912.00271 (2019).

[17] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.

[18] Fan, Deng-Ping, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. "Inf-Net: Automatic COVID-19 Lung Infection Segmentation from CT Images." IEEE Transactions on Medical Imaging, 2020.

[19] Zhou, Zongwei, Md Mahtfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. "Unet++: A nested u-net architecture for medical image segmentation." In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 3-11. Springer, Cham, 2018.

[20] Chen, Liang-Chieh, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. "Encoder-decoder with atrous separable convolution for semantic image segmentation." In Proceedings of the European conference on computer vision (ECCV), pp. 801-818. 2018.

[21] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431-3440. 2015.

[22] Elharrouss, Omar, Nandhini Subramanian, and Somaya Al-Maadeed. "An encoder-decoder-based method for COVID-19 lung infection segmentation." arXiv preprint arXiv:2007.00861, 2020.

[23] Ma, Jun, Yixin Wang, Xingle An, Cheng Ge, Ziqi Yu, Jianan Chen, Qiongjie Zhu et al. "Towards Efficient COVID-19 CT Annotation: A Benchmark for Lung and Infection Segmentation." arXiv preprint arXiv:2004.12537, 2020.

[24] Yu-Huan Wu, Shang-Hua Gao, Jie Mei, Jun Xu, Deng-Ping Fan, Rong-Guo Zhang, and Ming-Ming Cheng. "JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation." arXiv preprint arXiv: 2004.07054, 2020.

[25] He, K, Zhang, X, Ren, S, Sun, J, "Delving deep into rectifiers: Surpassing human level performance on imagenet classification." arXiv:1502.01852, 2015.

[26] A Chambolle, "An algorithm for total variation minimization and applications," Journal of Mathematical imaging and vision 20.1-2: 89-97, 2004.

[27] Minaee, Shervin, and Yao Wang. "An ADMM approach to masked signal decomposition using subspace representation." IEEE Transactions on Image Processing 28.7 (2019): 3192-3204.

[28] Zhang, Jian, Debin Zhao, and Wen Gao. "Group-based sparse representation for image restoration." IEEE Transactions on Image Processing 23.8: 3336-3351, 2014.

[29] https://github.com/DengPingFan/Inf-Net