Self-supervised Deep Learning Model for COVID-19 CT Lung Image Segmentation Highlighting Putative Causal Relationship Among Age, Underlying Disease and COVID-19

Daryl Fung Lerh Xing, Qian Liu, Judah Zammit, Carson Kai-Sang Leung, PingZhao Hu
University of Manitoba, MB, Canada

ABSTRACT COVID-19 is the new outbreak of a contagious disease that infects the lungs. Currently, no vaccines or antiviral medicines exist for COVID-19. As COVID-19 is a very contagious disease, cases appear faster than the amount of test kit available. Currently, the most common testing used is Polymerase Chain Reaction (PCR) test. These test would take several days for the test results to be available. Due to the exponential rate of infections, the limited amount of test kits, many infected patients are unable to get tested and receive treatments. An alternative approach to test for COVID-19 patients is through computerized tomography (CT) scan of the lungs. CT scan can drastically reduce the time taken for test results to be available and this could speed up the testing time as well as the limiting number of testing kits available. We propose a deep learning architecture that can evaluate different segmentation of the lungs from CT images to detect if a patient is infected with COVID-19 so that we can reduce the amount of time taken to carry out testing to determine if patients are infected with COVID-19. We use the publicly available CT lung images. We extend the work of InfNet and integrate self-supervised learning into InfNet to show that there is a performance improve. Our self-supervised InfNet approach obtained further improvement when we apply focal loss and lookahead optimizer. The improved performance can help prevent negatively assessing healthy patients which could prevent the patients from receiving unnecessary treatments that could cause negative side effects.

INDEX TERMS COVID-19, Self-Supervised Learning, Deep Learning, Lung CT Images, Image Segmentation

I. INTRODUCTION

COVID-19 is a newly identified disease that is very contagious and has been rapidly spreading across different countries around the world. The virus that was first identified in Wuhan has now infected more than 3.5 million people around the whole world and causes more than 245,000 deaths. Common symptoms from COVID-19 are fever, dry cough, but in more serious cases, patients can experience difficulty in breathing. As more people are infected, communities that have been in close contact with infected patients are getting tested for COVID-19. The test used to carry out the test for COVID-19 uses Polymerase Chain Reaction (PCR) test which could take several days for the test results to be available as the test samples are sent to a centralized lab for analysis and can be time-consuming. There is a limited number of supplies of PCR tests which is a bottleneck for testing to be efficient. Several alternative methods have been considered to test patients that are COVID-19 positive including a computerized tomography (CT) scan of the lungs. CT scans of the lungs are faster and easier to detect COVID-19 presence in patients. As the number of infected patients increases exponentially, it can be hard to provide testing scans for patients because of the limited number of doctors. It is recommended that Artificial Intelligence systems are used to analyze the CT scans of lung patients to determine the infected region of the lungs with COVID-19 and monitor the disease progression as well as to compensate for the high number of patients. Convolutional Neural Network (CNN) [1] is an important technique to be used in image processing as CNN is able to automatically captures useful features instead of handcrafting features to be used to evaluate on the segmentation of the CT lung images. Fan et al. [2] developed InfNet that uses CNN and fully supervised method to predict the segmentation of ground-glass opacities and consolidation. Fan et al. also incorporated semi-supervised learning to enlarge the limited number of training samples for CT lung image segmentation. However, deep learning requires a large of number of samples to be able to achieve good performance. Fan et al. uses pseudo labelling as semi-supervised learning to migitate the limited number of samples but pseudo labelling can take up to 3 times the time taken to train the network when undergoing fully supervised learning. Therefore, we propose using self-supervised deep learning to analyze and create a pixel-level segmentation of CT scan images of patients' lungs to determine the infected area of the CT lung images that includes ground-glass opacities and consolidation. Our self-supervised learning method integrated into InfNet is less time consuming than pseudo labelling as pseudo labelling requires training InfNet on the labeled data,

evaluating InfNet on the unlabeled data, then re-trained InfNet on the total dataset. Our *key contribution* in this paper is to integrate self-supervision into an existing network to improve the performance of the original network. We extend the work of InfNet as InfNet is one of the high performing model that includes CNN and several techniques to segment the ground-glass opacities and the consolidation area of the CT lung images. We show that integration self-supervised learning to InfNet improves the performance of InfNet.

II. Related works

Several works have been proposed to create image segmentation for CT scan lung images of COVID-19 positive patients. They have demonstrated effective solutions using deep neural networks to accurately predict if a patient has COVID-19 positive or negative.

A study has been conducted that uses supervised learning to train multiple models for different tasks where the study uses both classification and image segmentation tasks for COVID-19 detection through multi-tasks learning. The study uses Inception Residual Recurrent Neural Network (IRRCNN) for the classification of COVID-19 detection and uses Nabla-Net (NABLA-N) network for infected region segmentation for X-ray and CT images scan [3]. Transfer learning is used to retrain the IRRCNN model with samples to differentiate between COVID-19 positive samples and negative samples in the classification phase. Mathematical Morphological approaches are implemented for selecting appropriate contours for chest region selection in the segmentation phase with NABLA-N network. Some classical imaging and adaptive threshold approaches are applied to extract the features to identify infected regions of COVID-19. They used a total number of 5,216 samples of which 3,875 samples are pneumonia and 1,341 samples are normal.

Another study [4] introduces a supervised learning feature variation block and progressive atrious spatial pyramid pooling block using COVID-segNet, a high accuracy network that can create a segmentation of COVID-19 infection from chest CT images. The network consists of an Encoder and a Decoder with residual skip connection connecting the encoder and the decoder at their respective layer, following the architecture of UNET [5]. Their main findings include the introduction of an FV block and a PASPP block. FV block consists of three branches - contrast enhancement branch, position sensitive branch, and identity branch. These branches can enable automatic change of parameters to display positions and boundaries of COVID-19. The PASPP block takes features extracted from the FV block to acquire semantic information with a variety of receptive fields. The dataset that they used consists of 21,658 labeled chest CT images, of which 861 CT images are confirmed COVID-19.

The paper above however uses supervised learning and conducted the study with a good amount of data samples to train the network to achieve high performance. They obtained their dataset from hospitals through obtaining permission. We would like to create a network that does not require a much-labeled dataset to be able to achieve good performance. By doing this method, we could bring this network forward to detect new lung diseases when there is not many datasets available. Besides that, the paper is only able to recognize the presence of COVID-19 in a patient, but the papers could not quantify the severity of the disease.

While there is a limited number of public data samples available for CT COVID-19 lung image segmentation, it will not be feasible to train a network to achieve high performance. As there are not many COVID CT dataset that contains the segmentation ground-truth, we train a network that contains the segmentation of the infected region so that the prediction results from our model are more intuitive and easily comprehensible. Several kinds of research that resolve this issue. One method is to use semi-supervised learning to mitigate the problem of having a low number of data samples to improve the performance of deep neural networks. Instead of having to manually annotate the data, semi-supervised learning utilizes the unlabeled data samples to aid in the training for the network.

Fan et al. [2] used semi-supervised learning to enlarge the limited number of training samples for CT lung image segmentation. They developed a model called InfNet and semi-InfNet. The InfNet version of the model uses a fully supervised method to predict the segmentation of the CT images for ground-glass opacities and consolidations. The model outputs 4 images of the segmentation for the CT lung images that contain either ground-glass opacities or consolidations with different image sizes. The segmentation of the different image sizes is resized to the same size as the ground truth of the segmentation to compute the loss function. They also use an edge loss to guide the model to predict the boundary area of the segmentation. To improve InfNet, they use semi-supervised by progressively enlarging the training dataset with unlabeled data using a random sampling strategy. Specifically, they generate pseudo labels for unlabeled CT lung images. The advantage of using semi-supervised learning is that we can generate pseudo labels to increase the number of data samples. However, semi-supervised learning still requires to generate new examples through the use of unlabeled CT lung images before being able to undergo its learning procedure. This requires the use of unlabeled CT lung images to generate weakly labeled samples that are treated normally as labeled CT lung images to be fed into the network to train. This would be more time consuming as the network would have to first be trained on the labeled CT lung images, then evaluate the trained network on the unlabeled CT lung images to convert the unlabeled CT lung images into labeled CT lung images. After which the whole labeled CT lung images would be retrained again. This would take more than 3 times the time to train a supervised version of the network.

Another study [6] uses Task-Based Feature Extraction Network (TFEN) and Covid-19 Identification Network (CIN). They propose to use a task-specific feature extraction network that is tailored to CT lung images with three different classes: Healthy, pneumonia, and COVID-19 cases. They also mentioned that the dataset for COVID-19 is still limited and there is not enough high-quality dataset. They treat the task-

specific feature extraction network as autoencoders and train the overall TFEN module to extract the relevant features from the CT images. Then, they use CIN to perform classification on the extracted features from the TFEN module. They can easily detect the abnormal regions and differentiate between them very accurately by making use of prior information even when a person contains limited CT images. This helped them develop a semi-supervised feature extraction network that allows obtaining the relevant prior information to perform the classification to mimic human behaviors. However, this study does not undergo segmentation of the CT lung images for better diagnosis of the CT lung images.

There is a study that predicts the severity score of COVID-19 on chest x-ray with deep learning [7]. They use a DenseNet model from the TorchXRayVision library as DenseNet models have been shown to predict Pneumonia well. They use a pre-training step to train the feature extraction layers and a task prediction layer. The pre-training step was used to generate general representations of lungs and other Chest X-rays (CXRs) that they would have unable to achieve from the small set of COVID-19 images available. They use a network that outputs 18 outputs of a representation of the image, 4 outputs that are a hand-picked subset which contains the radiological findings (pneumonia, consolidation, lung opacity, and infiltration), and a lung opacity output. This study however did not use evaluate the segmentation of the infected region. They were only able to classify the CT lung images.

Lin et al. [8] noticed that class imbalance encountered during the training of dense detectors tend to overwhelm the cross entropy loss function. The negative samples that are easily classified comprise the majority of the loss and have a huge influence on the gradient. Instead, they propose a loss function, focal loss, to reduce the weight of easy examples and focus more on hard negatives. Focal loss was able to improve the performance of the dense detector when the dataset trained on has class imbalance.

SGD remains a popular optimizer to be used in training deep learning networks. There are improvement to SGD that includes AdaGrad [9] or Adam [10] which uses adaptive learning to optimize the weights of the deep learning network. However, hyperparameter tuning are costly to ensure the improvement of performance in deep learning networks with adaptive learning. Zhang et al. [11] present Lookahead that is less sensitive to suboptimals hyperparameters and reduce the need for additional effort to tune the hyperparameter. Lookahead optimizer warps around SGD or Adam and is able to achieve fast convergence across multiple deep learning tasks with minimal computational overhead.

III. Problem Statements

There are several limitations in the related works described. First, getting a high performance in deep neural networks requires an abundant amount of annotated samples. Performance can be drastically reduced if there are not enough data samples to compensate for the model's complexity. Second, learning complex data distributions require a higher model complexity to be able to fit the distribution with better performance. The related works utilize semi-supervised learning to increase the number of data samples to achieve higher performance. As pixel-level segmentation on CT images is a complex task, pixel-level segmentation requires a high model complexity to fit the distribution. Unfortunately, there is a limited number of publicly available COVID-19 datasets especially in the form of pixel-level segmentation. The limited number of samples available greatly reduces the performance of modeling complex distribution for pixel-level segmentation of CT scans lung images.

To solve the challenges, We propose a model and technique that utilizes self-supervised learning to mitigate the limited number of publicly available COVID-19 CT lung images samples to segment the infected regions of CT lung images.

IV. Methodology

In this section, we show the details of the self-supervised InfNet for imaging segmentation model including the network architecture, the data preprocessing steps, and the loss function. We show how self-supervised InfNet helps to improve generalization and performance of the model while having a limited number of data samples. We also show the extension of our data preprocessing steps.

Supervised InfNet (Lung Infection Segmentation Network) is used as our baseline to compare without using any semi-supervised learning algorithm. This is to show that the self-supervised learning method improves the performance of the baseline supervised learning InfNet for imaging segmentation. We extend our work on supervised InfNet by adding self-supervision method to it.

We do not change the structure of the InfNet model and use the default parameters as included in their GitHub code. There are two different types of the InfNet model - single InfNet and multi InfNet.

The output of the single segmentation InfNet includes the edge of the segmentation and four single-labeled segmentation of the infected region of the CT lung images with different sizes as shown in Figure 1. The single InfNet creates a single-labeled segmentation of the image for the infected region Figure 2A. The single InfNet predicts if the region is either ground-glass opacities or consolidations. It represents ground-glass opacities or consolidations as the same label. This means that the single InfNet will only predict the infected region without classifying them more specifically. The CT lung image is first passed into the initial convolutional layers of the single InfNet to extract the features of the CT lung image. Then, the features generated from the convolutional layer are fed into the partial decoder module, reverse attention module, and the edge detection module. The edge detection module is to help the network with the detection of the boundaries of the segmentation. The reverse attention and the partial decoder generates the segmentation of the infection regions of the CT lung images.

The prediction from the single InfNet represents the infected region and act as a prior to be fed into the multi InfNet Figure 3A. The prior is concatenated with the original CT image to be fed into the multi InfNet network. The multi

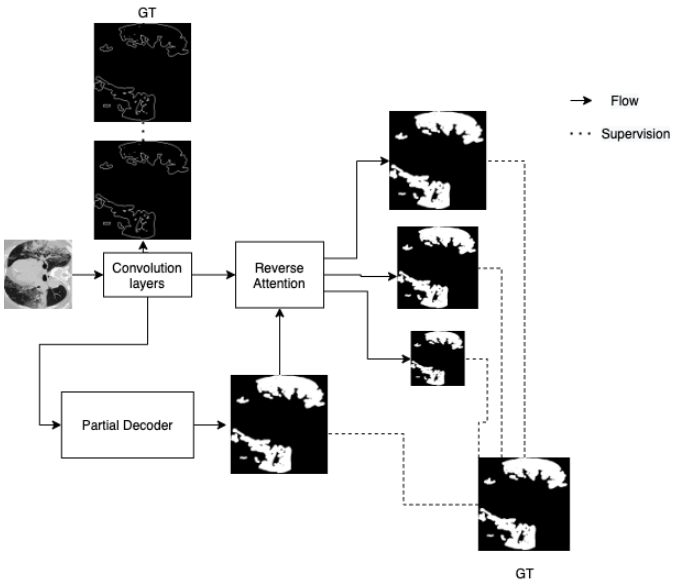

Fig. 1. Architecture of the supervised InfNet.

InfNet network is used to predict multiple-labeled segmentation. The multiple-labeled segmentation includes predicting the background, ground-glass opacities, and consolidations for the infected region. The multiple-labeled segmentation model gives each of the labels a different value instead of grouping them as one as what the single-segmentation model does.

A. *Self-supervised InfNet for imaging segmentation*

We propose using a self-supervised method to improve the performance of deep neural networks to create pixel-level segmentation for CT scans for lung images of COVID-19 patients. We integrate self-supervised inpainting to pre-train our network. Since image inpainting is similarly related to image segmentation, we integrate the pre-training steps as image inpainting for our image segmentation network.

The original InfNet model would generate 5 different predictions: the edge segmentation prediction and the other 4 are segmentation of the infected regions but of different sizes. To utilize the ability of self-supervised method for InfNet segmentation, we generate masks to be fed into the InfNet model. The last convolution layer that outputs the prediction is not used for the self-supervised case. However, the last convolutional layer is replaced with a different convolutional layer to reconstruct the image and the edge appropriately. Everything else is kept the same as the InfNet architecture — Figure 2B. This way the network learns meaningful representations of the CT images and we can use these meaningful representations to learn the segmentation of the infected regions of the CT lung images. After learning the self-supervised features for InfNet, the training continues as normal similar to the InfNet algorithm. The training starts with the weights trained using the self-supervised inpainting method. The last layer is changed to its original layer instead of the replaced convolutional layer.

By learning features from image inpainting, the model can learn more features that are related to image segmentation. As creating masks can be a complex task for the network to learn to inpaint, the mask can either be too complex for the network to start learning or too simple to be able to learn good representations. We use a coach network that increases the complexity of the masking of the CT images throughout the training of the network. The mask created is initially simple, once the network can predict the inpainting of the CT images with good performance, the coach increases the complexity of the masking to reduce the performance of the network, similar to how Generative Adversarial Network (GAN) works. The loss for the coach network is constructed from the loss of the image inpainting from the InfNet. The coach network and the InfNet both work together as a MinMax algorithm. The InfNet tries to minimize the loss to generate better image inpainting while the coach network tries to increase the loss of the image inpainting through generating more complex masks. In the beginning, the masks generated by the coach network is less complex. Through the training of the coach network, as the InfNet gets better at predicting image inpainting, the coach network starts to generate more complex masks. The loss fuction for the coach network is:

$$L_{coach}(x) = 1 - L_{rec}(x \odot M) \tag{1}$$

where

- M is the mask created by the coach network
- x is the CT lung image
- L_{coach} is the loss for the coach network
- L_{rec} is the loss for the reconstruction loss

A constraint is applied to this loss function because the coach network would just create a mask that masks all regions. After all, no context information would be present for the network to learn and a maximum loss is achieved. The constraint is:

$$\hat{B}(x) = B(x) - SORT(B(x))^{k|B(x)} \tag{2}$$

$$M = C(x) = \sigma(\alpha\hat{B}(x)) \tag{3}$$

The backbone, B, of the coach network has a similar network architecture with the model that inpaints the CT images. SORT(B(x)) sorts the features in descending order over the activation map. k represents the k^{th} elements in the sorted list and k helps to control the fraction of the image to be erased. The regions that has scores lesser than the k^{th} element are erased from the images. If k is 0.75 then 0.75 fraction of the images is not erased. The score is scaled into a range of [0, 1] using a sigmoid, σ, activation function. $C(x)$ is the coach network that is fed with the CT lung images. We keep $\alpha = 1$ while training the coach network. The illustration of the coach network can be seen in Figure 4.

After the self-supervision training is finished, the single segmentation InfNet would reuse the self-supervised single InfNet network weights to train normally on the segmentation of the CT lung images. Likewise, the multi InfNet network would reuse the weights that were trained during self-supervised multi InfNet training to train normally on the segmentation of the CT lung images.

The proposed self-supervised single-labeled segmentation InfNet network architecture can be seen in Figure 2B. The left side of the figure is the original Single InfNet architecture and the right side of the figure is the self-supervised Single InfNet.

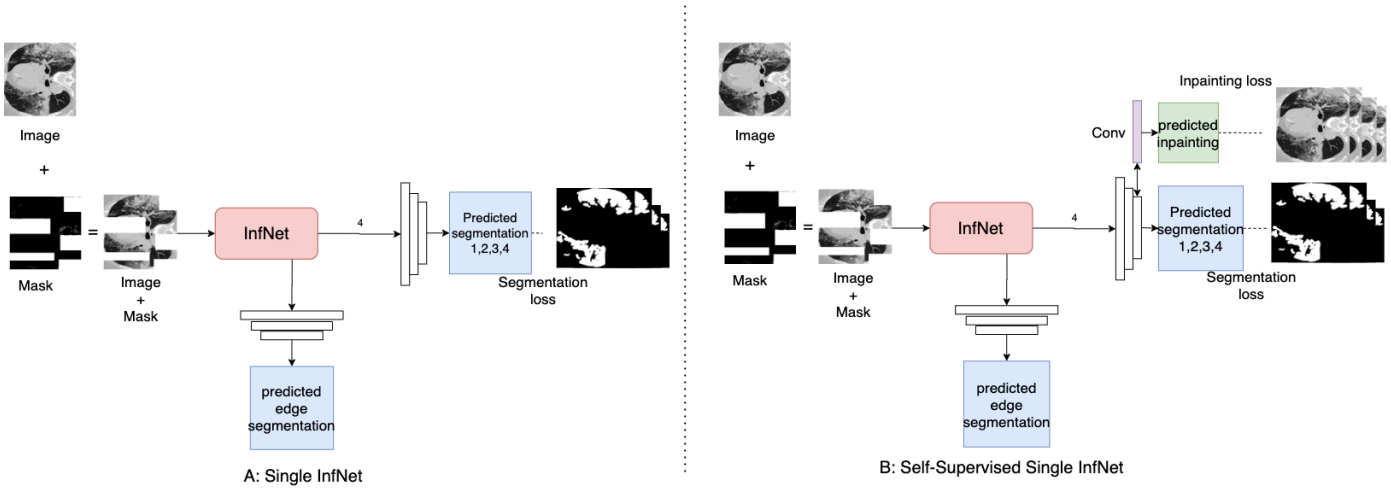

Fig. 2. A is the original architecture of the InfNet model. B is the architecture of our self-supervised InfNet model. Highlighted purple block is the difference between the original Single InfNet and the self-supervised Single InfNet.

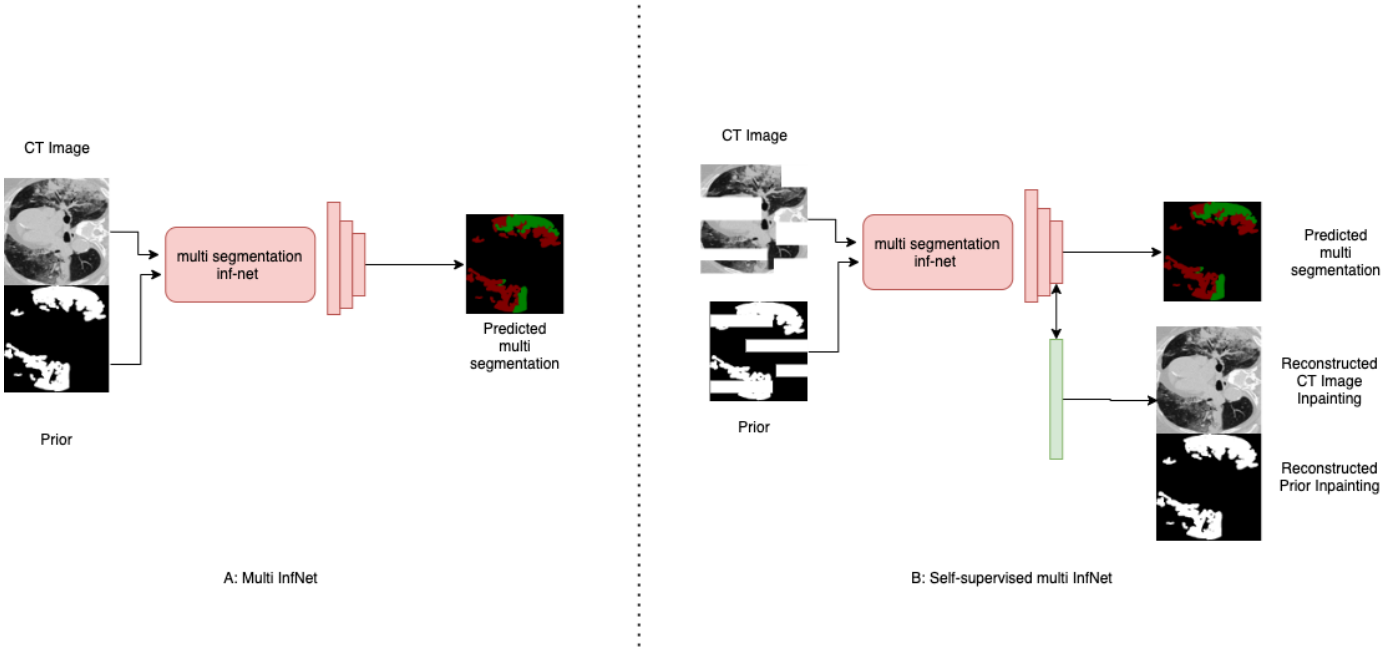

Fig. 3. A is the architecture of the original multi segmentation InfNet model. B is the architecture of our self-supervised multi segmentation InfNet model. Highlighted green block is the difference between the original multi InfNet and our self-supervised multi InfNet.

The last layer for each output prediction is replaced with a different linear activation layer. The linear activation layer re-creates the original image that is covered by the masks.

The proposed self-supervised multi-labeled segmentation InfNet network architecture is shown in Figure 3B. The changes in the architecture for the multi-labeled segmentation InfNet are similar to the single-labeled segmentation InfNet where the last layer of the layer is replaced with a different linear activation layer to output the inpainting of the original image.

A loss is calculated for each of the outputs from the single InfNet model. The first loss function is the loss edge, L_{edge} which guides the model in representing better segmentation boundaries. The other loss function is the segmentation loss, L_{seg}. The segmentation loss combines both the loss of Intersection over Union (IoU) and the binary cross entropy loss. The segmentation loss equation for the single InfNet is as follow:

$$L_{seg} = L_{IoU} + \lambda L_{BCE} \tag{4}$$

λ is a hyperparameter that controls how much weight we want to emphasize on the binary cross entropy loss. The λ is set to 1 for this experiment. The segmentation loss is adapted to all of the S_i predicted output where S_i are created from f_i such that $i = 3,4,5$.

The total loss function for the single InfNet model is then:

$$L_{total} = L_{seg}(G_t, S_g) + L_{edge} + \sum_{i=3}^{5} L_{seg}(G_t, S_i) \tag{5}$$

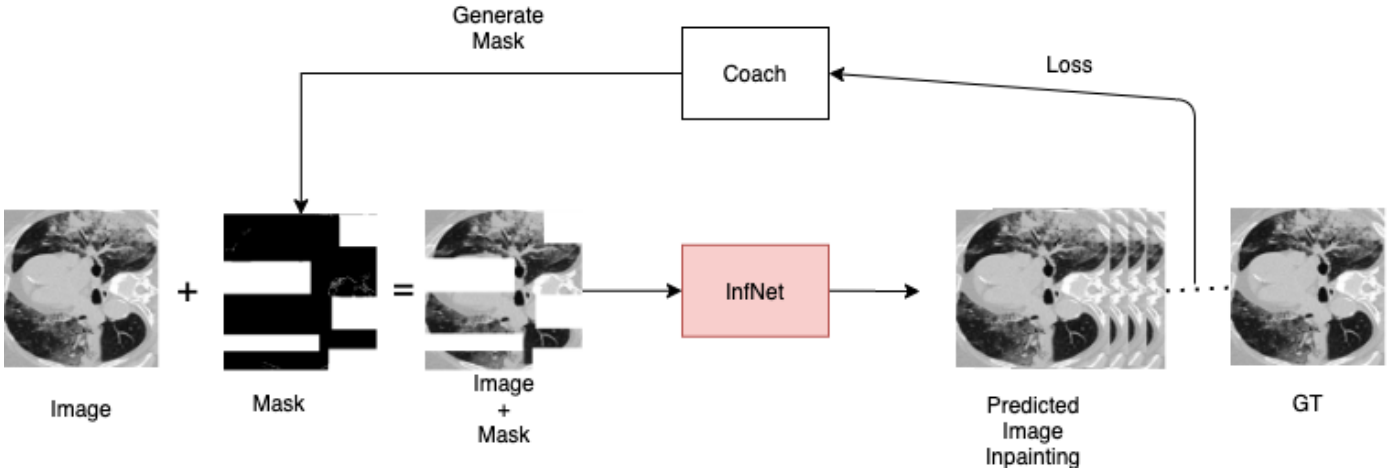

Fig. 4. The architecture of the coach network for self-supervised inpainting.

The summation of the segmentation loss functions are calculated from the output of the parallel partial decoder and the output of the three convolutional layers — 3rd layer to the 5th layer. G_t refers to the ground truth labels. S_g is the output from the parallel partial decoder to match with the ground truth label. S_i is the different sizes of the segmentation of infected region output by the InfNet. The different sizes of the segmentation of infected region outputted by the InfNet are resized to fit the same shape as the ground truth segmentation image. L_{edge} is the loss for the edge.

As for the multiple segmentation infected region InfNet. We also use the default model and hyperparameters from the InfNet code. However, we train the network without using any unlabeled images to be used as a supervised version. The CT lung images and prior (infected region) for the CT lung images are concatenated together before being fed into the multiple segmentation InfNet. The prior is generated from the single segmentation InfNet. The prior would contain the area of the infected region. However, the prior does not contain the labels for ground-glass opacities and consolidations. It just shows the infected regions. The multiple segmentation InfNet labels the CT lung images with background, ground-glass opacities, and consolidations. The architecture for multiple segmentation InfNet can be seen in Figure 3. The loss function for the multiple segmentation InfNet is as follow:

$$L_{bce} = \frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(\hat{y}_i) + (1-y_i)\cdot log(1-\hat{y}_i) \tag{6}$$

where

- y_i is the ground truth value for the segmentation - background, ground-glass opacities, and consolidation
- $\hat{y}_i$ is the predicted value for the segmentation by the network
- N is the total number of the current training batch of data samples

The loss function for multiple segmentation InfNet uses the binary cross-entropy between the predicted segmentation and the ground truth segmentation. 365 In order to improve the performance of the model and to aid in the generalization, we determine to use self-supervised learning to learn good representations of the CT scan of lung images. Self-supervised learning generates auxiliary tasks from the labeled data samples. For instance, we could train the network to predict if the images have been rotated 0°, 90 0°, 180 0° to learn representations of the images.

Additionally, we use focal loss instead of the binary cross entropy loss function for the self-supervised InfNet model to emphasize on the smaller data samples on consolidation than ground-glass opacities. The focal loss function is:

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma} log(p_t) \tag{7}$$

where

- FL is the focal loss
- p_t is the predicted output by the Multi InfNet
- α_t is a hyperparameter that controls the weight of positive and negative samples
- γ is the term that controls the rate of the downweighted examples

We also wrap Lookahead optimizer around the SGD optimiser with k = 5 and alpha = 0.5.

The pseudo code for our self-supervised Single/Multi InfNet can be seen in Algorithm 1.

V. Experiments

A. Datasets

The dataset that we use is an integrative resource of chest computed tomography images and clinical features of patients with COVID-19 pneumonia (ICTCF) [12] which contains the severity score for each CT lung image and CT lung images from medical segmentation website [13].

ICTCF contains 127 types of clinical features and laboratory-confirmed cases of COVID-19 from 1170 patients including the severity of the CT lung images. However, the ICTCF dataset does not contain the segmentation labels for the ground-glass opacities and the consolidation in the CT lung images. In total, there are 6654 of CT lung images in ICTCF dataset. Originally, there were 1521 patients. However, some of the patients are missing CT lung images. We remove these

Algorithm 1 Pseudo code for self-supervised with InfNet

Input: $D_{labeled}$ = [($inputImage_1$, G_{t1}), ...]
for each epoch **do**
  for each coach step **do**
    mask = M(x)
    maskedInput = $mask \odot inputImage$
    $predictedImage = network(maskedInput), inputImage$
    $L_{rec} = CrossEntropy(predictedImage, inputImage)$
    $L_{coach}(x) = 1 - L_{rec}$
    update coach weights
  end for
  for each network step **do**
    $P_{labeled} = Preprocess(D_{labeled})$
    $inpaintingOutput = network(P_{labeled})$
    $L_{rec} = CrossEntropy(InpaintingOutput, inputImage)$
    backpropogate and save network weights
  end for
end for
for each batch of $D_{labeled}$: **do**
  $P_{labeled}$ = Preprocess ($D_{labeled}$)
  $trainLoss = train(P_{labeled})$
  Backpropagate train loss
  $testLoss = test(P_{labeled})$
  save model weights, w.
end for

patients that are missing CT lung images. After preprocessing the patients, the dataset was left with 1338 patients that contain CT lung images. The dataset can be found here: http://ictcf.biocuckoo.cn/. We use these ICTCF CT lung images without the ground truth segmentation labels in combination with the MedSeg dataset to undergo self-supervised learning to predict image in-painting.

As for the MedSeg dataset, they contain ground truth labels for the segmentation for ground-glass opacities and consolidation of the CT lung images. The total amount of CT lung images contain in MedSeg dataset is 932 CT lung images. We randomly assign the CT lung images into a training set, validation set, and testing set of which the training set contains 698 CT lung images, the validation set contains 114 CT lung images, and the testing set contains 117 CT lung images.

The assignment of the dataset can be seen in Table I.

Data split	Source	Segmented	Images	Patients
Training	Med-Seg	Yes	698	39
	ICTCF	No	6654	1338
Validation	Med-Seg	Yes	114	35
Testing	Med-Seg	Yes	117	35

TABLE I. This table shows the data distribution between the datasets that we use to evaluate our model on. Med-Seg refers to the COVID-19 CT Segmentation data set and ICTCF refers to the ICTCF data set.

B. Experimental Settings

During the self-supervised image inpainting stage, we train the network for 2000 epochs. The network is trained for the first 200 epochs before we train the coach network for 200 epochs which increases the complexity of the masks generated. After that, we alternate in between training the self-supervised image inpainting and the coach network with 100 epochs in between. For every alternating between the training of the self-supervised image inpainting and the coach network, we set the learning rate to 0.1 at the start of the epoch, we set the learning rate to 0.01 at 40th epoch, we set the learning rate to 0.001 at 80th epochs, and 0.0001 at the 90th epoch. We use SGD as the optimizer for the self-supervised image inpainting. We set the momentum to 0.9 and the weight decay to 0.0005. As for the optimizer for the coach network, we use Adam optimizer with a learning rate of 0.00001.

For the Single InfNet, we train the network for 500 epochs. We use Adam as the optimizer with a learning rate of 0.0001.

For the Multi InfNet, we train the network for 500 epochs. We use SGD as the optimizer. The momentum is set as 0.7 and the learning rate is set as 0.01.

As for the Multi InfNet with added focal loss and lookahead optimizer, we trained the network for 500 epochs, we use lookahead optimizer with k=5 and alpha=0.5 and warp the lookahead optimizer around SGD optimizer where the momentum is set as 0.7 and the learning rate is set as 0.01.

For the self-supervised version for both Single InfNet and Multi InfNet, the self-supervised image in-painting is first trained. Then the weights from the trained networks except for the last layer are transferred to be used to train on the segmentation of the CT lung images.

We compare our method against the supervised [14] models trained on COVID-19 dataset. We train and follow using the same network structure but change from supervised learning to self-supervised learning and compare the performance between supervised and self-supervised.

We use this approach to determine if self-supervised learning can be a useful task to help InfNet improve its performance in segmenting the ground-glass opacities or consolidation around the infected region of the CT lung images.

C. Performance Evaluation Metrics

we will show the results of our experiments obtained. We show the comparison of the results between the supervised and the self-supervised version of the InfNet.

The table is plotted with several metrics: F1, IoU, Recall, and Precision.

The F1-Score is also called the Dice Coefficient, it is used to measure the overlap between the ground-truth infected region and the predicted infected region. The F1-Score equation is defined as:

$$F1 = \frac{2 * |T \cap P|}{|T| + |P|} \tag{8}$$

where T is the ground truth infected region and P is the predicted infected region.

The IoU is a different method to measure the overlap between the ground truth infected region and the predicted infected region. The IoU equation is defined as:

$$IoU = \frac{T \cup P}{T \cap P} \tag{9}$$

where T is the ground truth infected region and P is the predicted infected region.

The Recall is used to measure the how much of the ground truth infected region is present in the predicted infected region. The equation is as follow:

$$Recall = \frac{T \cap P}{T} \tag{10}$$

where T is the ground truth infected region and P is the predicted infected region.

The Precision is used to measure how much of the predicted infected region is present in the ground truth infected region. The equation is as follow:

$$Precision = \frac{T \cap P}{P} \tag{11}$$

where T is the ground truth infected region and P is the predicted infected region.

There are cases when the calculated F1, IoU, Recall, or Precision will contain NaN value due to the fact that the denominator is 0. We will ignore NaN values into our calculation and continue the calculation for other metrics.

For the table that contains mean and error, the mean are calculated as:

$$mean = \frac{\sum_{i=1}^{N} Metric(\hat{y}_i, y_i)}{N} \tag{12}$$

Where Metric refers to either *F1, IoU, Recall, Precision.* N refers to the number of test data samples. The error is:

$$error = SE \times 1.96 \tag{13}$$

where SE is the standard error of the test data samples for the metric multiplied by 1.96. Note that Mean ±Error is the 95% confidence interval.

VI. Result

In this section, we show the results of our experiments obtained. To evaluate the performance of our proposed self-supervised InfNet, we compare its performance with other state-of-the-art methods including the original InfNet.

A. Result comparison for Single InfNet

Table IV shows the result for the single segmentation InfNet. The single segmentation InfNet does not segment between ground-glass opacities or consolidation. The single segmentation segments and represents all infected region as one. We can see that self-supervision can improve on the generalization and consistency in predicting the different CT lung images as they perform the best in terms of the error range. Even though the baseline single InfNet (Single SInfNet) performance has better mean values for F1, IoU, and Recall, the self-supervised approach helps to create robustness and consistency in the model itself to better handle outliers. We can see the results of the single segmentation in Figure 5. We can see that the baseline single InfNet (Single SInfNet) overestimated the infected region of an outlier in the segmentation result in the figure in the last row. The self-supervised Single InfNet (Single SSInfNet) did a better job at predicting outliers where its prediction is more closely related to the ground truth than the baseline single SInfNet (Single SInfNet). The receiving operating characteristics (ROC) can be seen in Figure 6. The baseline Single InfNet (Single SInfNet) has the best performing ROC.

B. Results comparison for Multi InfNet

Table V shows the result for the comparison between multiple segmentation InfNet. As the multiple segmentation InfNet requires a CT lung image concatenate with a prior as input where the prior is the segmentation of the infected region of the CT lung without considering the location of ground-glass opacities or consolidation. The prior represents the infected region as a whole. The prior is obtained by running prediction of the infected region by the single segmentation InfNet on the CT lung images of the test set. Then the prior is fed together with the CT lung image from the test set into the multiple segmentation InfNet to obtain the result. As the baseline Single InfNet (Single SInfNet) achieves the best performing single InfNet, we use the prediction of the prior obtained from the baseline Single InfNet to be fed into the multi segmentation InfNet with the CT lung images. The self-supervised Multi InfNet (SSInfNet) was able to achieve a higher performance than the baseline Multi InfNet (SInfNet). The self-supervised Multi InfNet (SSInfNet) achieves the best performance in evaluating the ground-glass opacities and the consolidation area of the CT lung images than the baseline Multi InfNet (SInfNet). The overall performance of the self-supervised Multi InfNet (SSInfNet) achieves the best result. We can see the segmentation result in Figure 7. However, the baseline Multi InfNet (SInfNet) achieves a better recall score than the rest of the networks. As we can see in Figure 7, the baseline multi InfNet (SInfNet) predicted more of the consolidation even on areas that is not infected. On the third row, the SInfNetoverestimated the consolidation area in the healthy CT lung image. As recall is the amount of true positive over the total actual consolidation area, the SInfNet seems to overestimated the consolidation area which results in a higher recall score than the other networks. However, as most of the prediction of the consolidation area for the SInfNet is not accurate, the precision for the SInfNet is lower. This therefore decreases the performance of the SInfNet.

VII. Conclusion

Our results show that the integration of self-supervised image in-painting to the supervised multi segmentation InfNet improves the performance of both the ground-glass opacities and the consolidation in the dataset that we used. Additionally, adding focal loss and lookahead optimizer further improves our self-supervised multi InfNet and achieves 0.63 F1 scores. The improvement in the performance of the infected region segmentation for ground-glass opacities and consolidation of CT lung images can help prevent negatively assessing that a patient contains irregular patterns when the patients are healthy. This

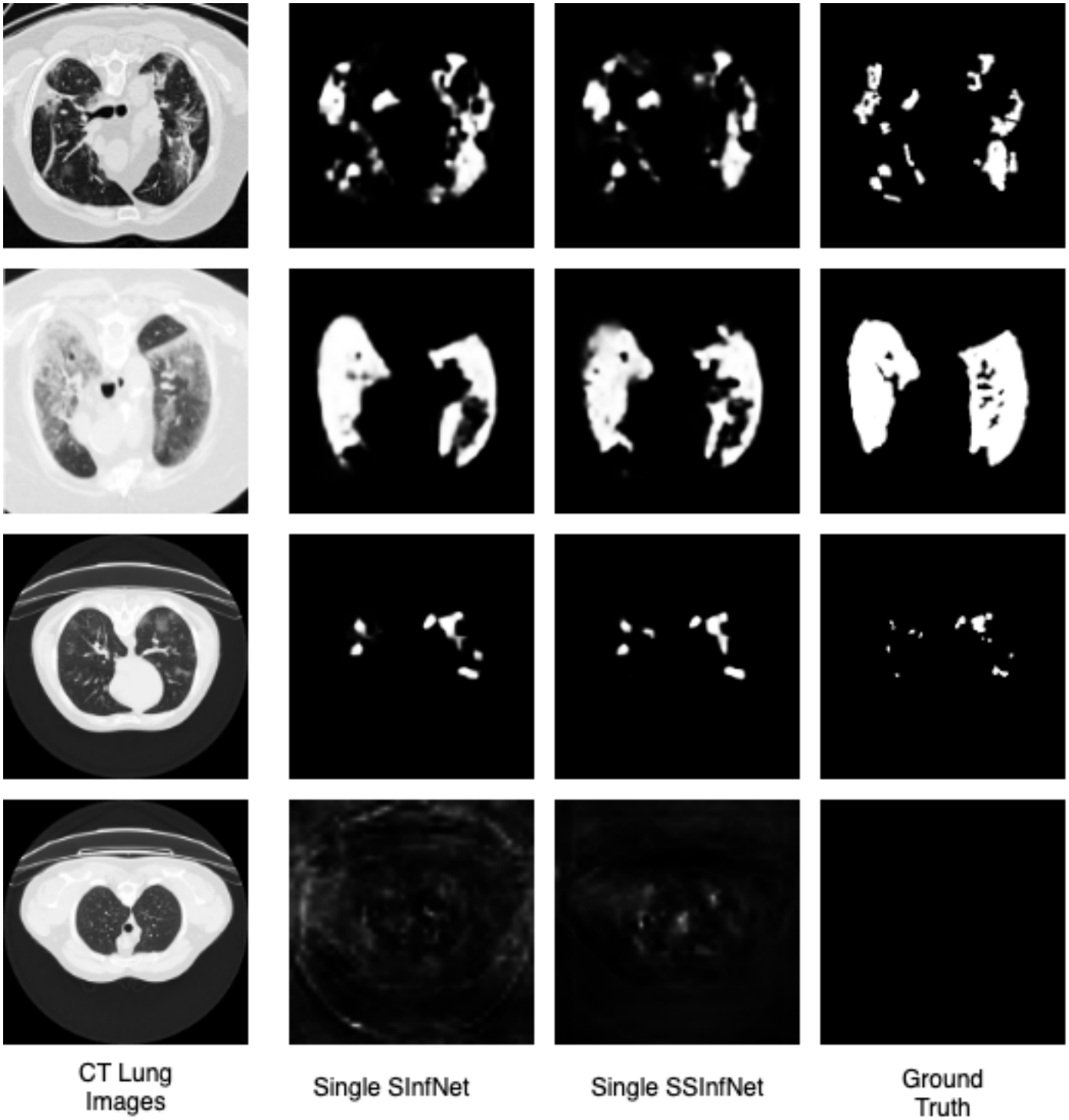

Fig. 5. Comparison of single segmentation between different networks. Overall, the Single SInfNet achieves a better performance than the rest of the network as seen in row 1 to row 3. However, our method by adding self-supervised to InfNet (Single SSInfNet) prevents overestimating the infected region of the CT lung images as can be seen in row 4.

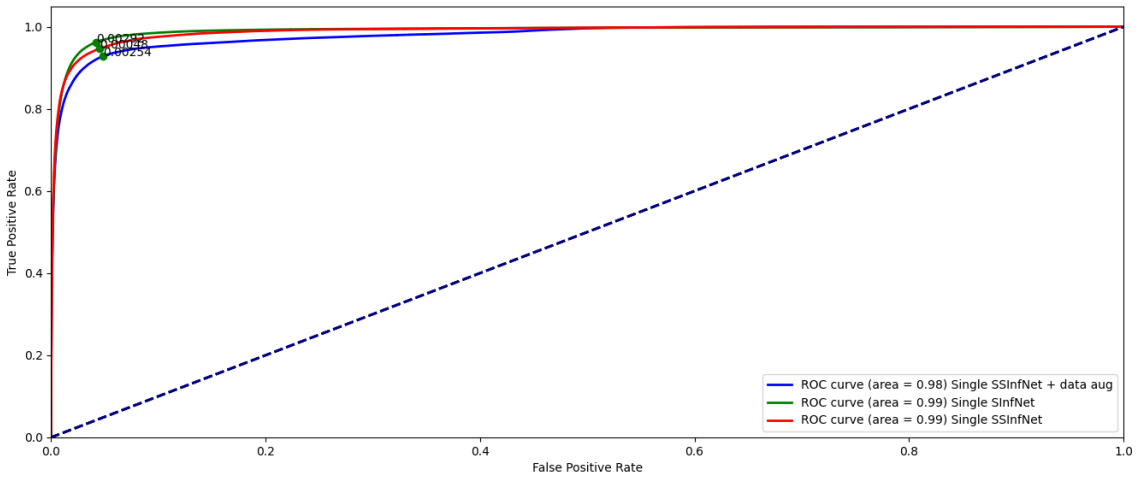

Fig. 6. ROC comparison of different networks.

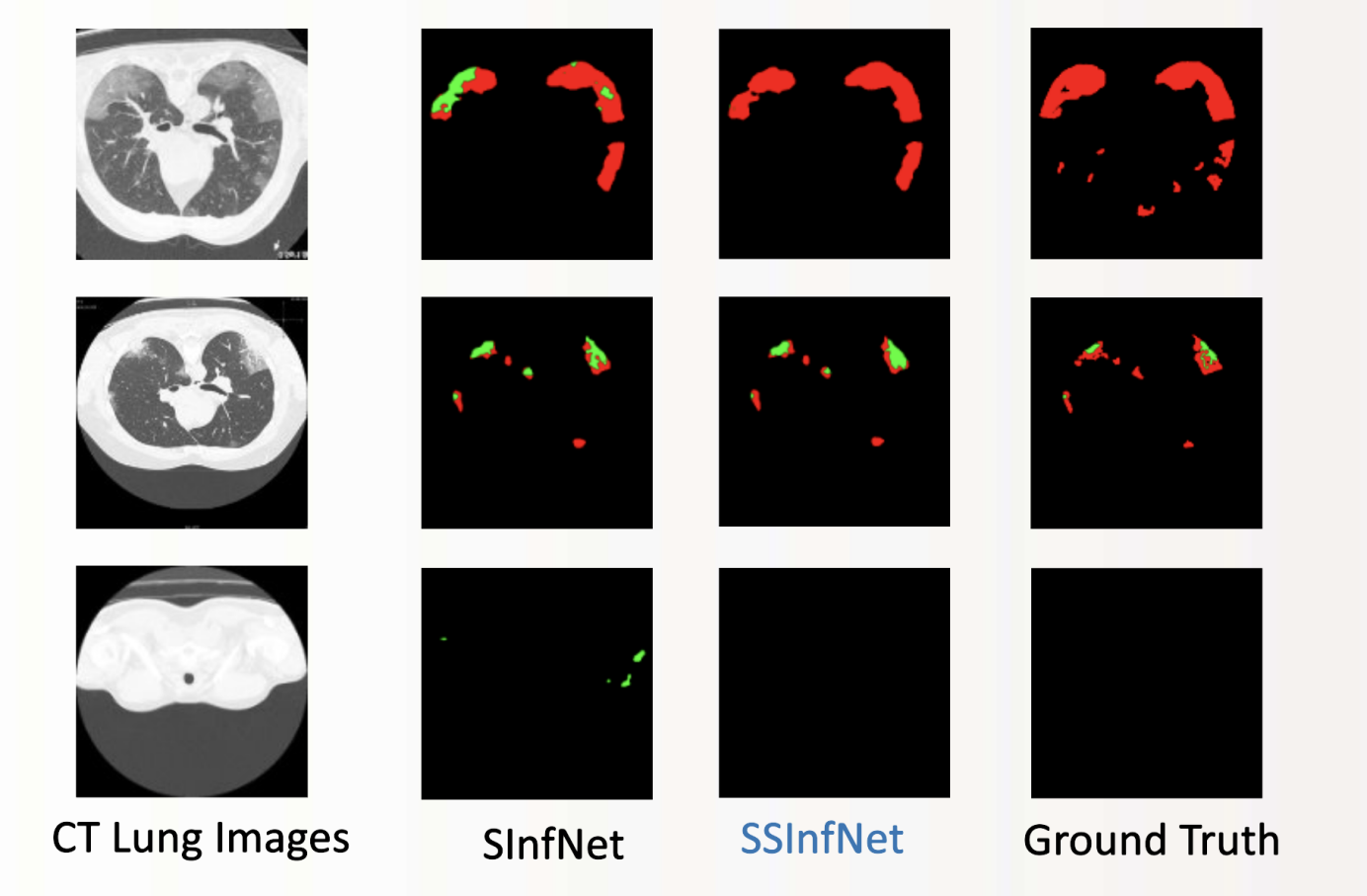

Fig. 7. Comparison of multi segmentation between different networks with prior generated from single InfNet. Red colored is the ground-glass opacities and green colored is the consolidation. The baseline multi InfNet (SInfNet) overestimated the consolidation region in the first and third row of the CT lung image when there are not suppose to be any consolidation region.

Methods		F1	IoU	Recall	Precision	AUC
Single SInfNet	Mean	**0.39**	**0.29**	**0.83**	0.33	**0.9909**
	Error	± 0.059	± 0.053	±0.069	±0.057	±0.032
Single SSInfNet	Mean	0.38	0.27	0.75	0.33	0.9883
	Error	±0.056	±0.049	±0.077	±0.053	± 0.010

TABLE II. Quantitative result for comparison between Single segmentation InfNet and self-supervised single segmentation InfNet in the test set. Single SSInfNet is our self-supervised single InfNet.

Methods		U-Net		SInfNet		SSInfNet	
		mean	error	mean	error	mean	error
GGO	F1	0.26	±0.057	0.38	±0.054	**0.43**	±0.057
	IoU	0.18	±0.043	0.27	±0.042	**0.31**	±0.046
	Recall	0.216	±0.053	**0.58**	±0.065	**0.58**	±0.072
	Precision	0.405	±0.085	0.41	±0.058	**0.48**	±0.059
Cons	F1	0.35	±0.097	0.29	±0.078	**0.46***	±0.096
	IoU	0.26	±0.08	0.22	±0.068	**0.36***	±0.088
	Recall	0.32	±0.089	**0.61**	±0.099	0.56	±0.11
	Precision	0.46	±0.116	0.31	±0.084	**0.56***	±0.101
Background	F1	0.857	±0.01	1.0	±0.002	1.0	±0.002
	IoU	0.754	±0.017	0.99	±0.003	0.99	±0.003
	Recall	0.998	±0.001	0.99	±0.002	0.99	±0.002
	Precision	0.755	±0.017	1.0	±0.002	1.0	±0.002
Overall	F1	0.49	±0.055	0.55	±0.044	**0.63***	±0.052
	IoU	0.40	±0.046	0.5	±0.038	**0.55**	±0.046
	Recall	0.51	±0.048	**0.73**	±0.055	0.71	±0.061
	Precision	0.54	±0.073	0.57	±0.048	**0.68***	±0.054

TABLE III. Quantitative result of Ground-glass Opacities & Consolidation on the test data set. Prior is obtained from the single segmentation InfNet. SSInfNet is our self-supervised multi InfNet. GGO: Ground-Glass Opacity. Cons: Consolidation. A p-value is calculated using t-test to compare the performance between the SSInfNet and the SInfNet. For values where the p-value is less than 0.05 and is significant, they are marked will *.

would prevent patients from receiving unnecessary treatments that could cause side effects. For the future work, we could apply this technique for other datasets that include segmenting Leukemia or breast cancer images.

ACKNOWLEDGMENT

I would like to acknowledge Dr. PingZhao Hu and Dr. Carson Leung for the supervision of this Honours Project. I have learned a lot about the technique and best practices to write a good paper. I would like to thank Judah Zammit and Qian Liu too, they have recommended me several methods and tools to improve on the work for this project. I would also like to acknowledge Dr. Ruppa Thulasiram for the opportunity to be able to take this course with Dr. PingZhao Hu and Dr. Carson Leung as my supervisor. I would not be able to get this much opportunity and learned all these without them. I would like to add ICTCF and MedSeg as an acknowledgement for providing the publicly available COVID-19 CT lung images dataset. The publicly available dataset has helped us been able to carry out this research.

REFERENCES

[1] Krizhevsky, A, Sutskever, I. , and Hinton G. *Imagenet classification with deep convolutional neural networks.* In NIPS, 2012.

[2] Fan, DP., Zhou, T., Ji, GP., et al. *Inf-Net: Automatic COVID-19 Lung Infection Segmentation from CT Scans.* arXiv preprint arXiv:2004.14133v2, 2020.

[3] Alom, MZ., Rahman, MMS., Nasrin, MS., Taha, TM., Asari, VK. *COVID-MTNet: COVID-19 Detection with Multi-Task Deep Learning Approaches.* arXiv:2004.03747, 2020.

[4] Yan, Q., Wang, B., Gong, D., et al. *COVID-19 Chest CT Image Segmentation – A Deep Convolutional Neural Network Solution.* arXiv:2004.10987, 2020.

[5] Ronneberger, O., Fischer, P., and Brox, T. *U-net: Convolutional networks for biomedical image segmentation.* In MICCAI, pages 234–241. Springer, 2015. 2

[6] Khobahi, S., Agarwal, C., Soltanalian, M. *CoroNet: A Deep Network Architecture for SemiSupervised Task-Based Identification of COVID-19 from Chest X-ray Images.* In medRxiv, 2020.

[7] Cohen, J. P., Dao, L., Morrison, P., Roth, K., Bengio, Y., Shen, B., Abbasi, A., Hoshmand-Kochi, M., Ghassemi, M., Li, H., Duong, T. Q. *Predicting covid19 pneumonia severity on chest x-ray with deep learning*, arXiv preprint arXiv:2005.11856.

[8] Lin, TY., Goyal, P., Girshick, R., He, K., and Dollar, P. *Focal loss for dense object detection.* arXiv preprint arXiv:1708.02002, 2017

[9] Duchi, J., Hazan, E., and Singer, Y. *Adaptive subgradient methods for online learning and stochastic optimization.* Journal of Machine Learning Research, 12(Jul):2121–2159, 2011.

[10] Kingma, D.P., and Ba, J. *Adam: A method for stochastic optimization.* arXiv preprint arXiv:1412.6980, 2014.

[11] Zhang, M.R., Lucas, J., Hinton, G., and Ba, J. *Lookahead optimizer: k steps forward, 1 step back.* arXiv preprint arXiv:1907.08610, 2019.

[12] Ning, Lei, WS., Yang SJ., et al. (2020). *iCTCF: an integrative resource of chest computed tomography images and clinical features of patients with COVID-19 pneumonia.* 10.21203/rs.3.rs-21834/v1.

[13] *COVID-19 CT segmentation dataset*. Retrieved from http://medicalsegmentation.com/covid19/.

[14] Yan, Q., Wang, B., Gong D., et al. *COVID-19 Chest CT Image Segmentation – A Deep Convolutional Neural Network Solution.* arXiv preprint arXiv:2004.10987, 2020.
[15] Kalluri, T., Varma, G., Chandraker, M., and Jawahar, CW. *Universal semi-supervised semantic segmentation.* CoRR, abs/1811.10323, 2018.
[16] Misra, I., and van der Maaten, L. *Self-supervised learning of pretext-invariant representations.* arXiv preprint arXiv:1912.01991, 2019.
[17] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. *A simple framework for contrastive learning of visual representations.* arXiv:2002.05709, 2020.
[18] Newell, A., Deng, J. *How Useful is Self-Supervised Pretraining for Visual Tasks?* arXiv:2003.14323, 2020.
[19] Novosel, J., Viswanath, P., and Arsenali, B. *Boosting Semantic Segmentation With Multi-Task Self-Supervised Learning for Autonomous Driving Applications.* In Proc. of NeurIPS - Workshops, pages 1–11, Vancouver, BC, Canada, Dec. 2019.
[20] Kahl, F. *"Fine-grained segmentation networks: Self-supervised segmentation for improved long-term visual localization,"* in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 31–41.
[21] Chang, YC., Yu, CJ., Chang, SC., et al. *Pulmonary sequelae in convalescent patients after severe acute respiratory syndrome: evaluation with thin-section CT.* Radiology 2005; 236(3):1067-1075.
[22] Yang, R., Li, X., Liu, H., Zhen, Y., Zhang, X., Xiong, Q., et al. *Chest CT Severity Score: An Imaging Tool for Assessing Severe COVID-19. Radiol Cardiothorac Imaging.* 2020;2(2):e200047.
[23] Shan, F., Gao, Y., Wang, J., Shi, W., Shi, N., Han, M., Xue, Z., and Shi, Y. *Lung Infection Quantification of COVID-19 in CT Images with Deep Learning.* arXiv preprint arXiv:2003.04655, 1-19, 2020.
[24] Kolesnikov, A., Zhai, XH., and Beyer, L. *Revisiting self-supervised visual representation learning.* In Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
[25] Trinh, TH., Luong, MT., and Le, QV. *Selfie: Self-supervised pretraining for image embedding.* arXiv preprint arXiv:1906.02940, 2019.
[26] Frinken, V., Zamora-Martinez, F., Espana-Boquera, S., Castro-Bleda, M. J., Fischer, A., and Bunke, H. (2012). *Long-short term memory neural networks language modeling for handwriting recognition.* In Pattern Recognition (ICPR), 2012 21st International Conference on, pages 701–704. IEEE.
[27] LeCun, Y., Haffner, P., Bottou, L., and Bengio, Y. *Object recognition with gradient-based learning.* In Shape, contour and grouping in computer vision, pages 319–345. 1999.
[28] Kingma, DP. and Welling, M. *Auto-Encoding Variational Bayes.* In The 2nd International Conference on Learning Representations (ICLR), 2013.
[29] Goodfellow IJ., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, AC., and Bengio, Y. *Generative adversarial nets.* In Proceedings of NIPS, pages 2672– 2680, 2014.
[30] Zhao, JY., Zhang, YC., He, XH., Xie, PT. *COVID-CT-Dataset: a CT scan dataset about COVID-19.* arXiv preprint arXiv: 2003.13865, 2020.
[31] Cohen, JP., Morrison, P., and Dao, L. *COVID-19 Image Data Collection.* arXiv preprint arXiv: 2003.11597, 2020. https://github.com/ieee8023/covid-chestxray-dataset.
[32] Zhang, K., Liu, XH., Shen, J., et al. *Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements and Prognosis of COVID-19 Pneumonia Using Computed Tomography.* DOI: 10.1016/j.cell.2020.04.045.
[33] Singh, S., Batra, A., Pang, G., Torresani, L., Basu, S., Paluri, M., and Jawahar, C. V. *Self-supervised feature learning for semantic segmentation of overhead imagery.* In BMVC, 2018.

SUPPLEMENTARY MATERIALS

Methods		F1	IoU	Recall	Precision	AUC
Single SInfNet	Mean	0.59	0.43	0.95	0.44	0.9891
	Error	± 0.062	± 0.066	±0.023	±0.069	±0.032
Single SSInfNet	Mean	0.50	0.35	0.94	0.35	0.9854
	Error	±0.0778	±0.074	±0.034	±0.074	±0.008

TABLE IV. Quantitative result for comparison between Single segmentation InfNet and self-supervised single segmentation InfNet in the validation set. Single SSInfNet is our self-supervised single InfNet.

Methods		U-Net		SInfNet		SSInfNet	
		mean	error	mean	error	mean	error
GGO	F1			0.04	±0.018	0.05	±0.02
	IoU			0.02	±0.011	0.03	±0.012
	Recall			0.11	±0.048	0.11	±0.048
	Precision			0.1	±0.042	0.12	±0.05
Cons	F1			0.01	±0.01	0.02	±0.014
	IoU			0.01	±0.005	0.01	±0.008
	Recall			0.05	±0.038	0.05	±0.036
	Precision			0.02	±0.015	0.03	±0.024
Background	F1			0.98	±0.006	0.98	±0.006
	IoU			0.96	±0.011	0.96	±0.01
	Recall			0.98	±0.008	0.98	±0.008
	Precision			0.98	±0.007	0.98	±0.007
Overall	F1			0.34	±0.011	0.35	±0.013
	IoU			0.33	±0.009	0.33	±0.01
	Recall			0.38	±0.031	0.38	±0.031
	Precision			0.36	±0.015	0.38	±0.027

TABLE V. Quantitative result of Ground-glass Opacities & Consolidation on the validation data set. Prior is obtained from the single segmentation InfNet. SSInfNet is our self-supervised multi InfNet. GGO: Ground-Glass Opacity. Cons: Consolidation.