A Mask-guided Attention Deep Learning Model for COVID-19 Diagnosis based on an Integrated CT Scan Images Database

Maede Maftouni [1], Bo Shen [2], Andrew Chung Chee Law [2], Niloofar Ayoobi Yazdi [2], and Zhenyu Kong [2]

[1]Virginia Tech
[2]Affiliation not available

November 1, 2023

Abstract

The global extent of COVID-19 mutations and the consequent depletion of hospital resources highlighted the necessity of effective computer-assisted medical diagnosis. COVID-19 detection mediated by deep learning models can help diagnose this highly contagious disease and lower infectivity and mortality rates. Computed tomography (CT) is the preferred imaging modality for building automatic COVID-19 screening and diagnosis models. It is well-known that the training set size significantly impacts the performance and generalization of deep learning models. However, accessing a large dataset of CT scan images from an emerging disease like COVID-19 is challenging. Therefore, data efficiency becomes a significant factor in choosing a learning model. To this end, we present a multi-task learning approach, namely, a mask-guided attention (MGA) classifier, to improve the generalization and data efficiency of COVID-19 classification on lung CT scan images.

The novelty of this method is compensating for the scarcity of data by employing more supervision with lesion masks, increasing the sensitivity of the model to COVID-19 manifestations, and helping both generalization and classification performance. Our proposed model achieves better overall performance than the single-task baseline and state-of-the-art models, as measured by various popular metrics. In our experiment with different percentages of data from our curated dataset, the classification performance gain from this multi-task learning approach is more significant for the smaller training sizes. Furthermore, experimental results demonstrate that our method enhances the focus on the lesions, as witnessed by both

attention and attribution maps, resulting in a more interpretable model.

A Mask-guided Attention Deep Learning Model for COVID-19 Diagnosis based on an Integrated CT Scan Images Database

Maede Maftouni[a], Bo Shen[a], Andrew Chung Chee Law[a], Zhenyu (James) Kong[a]*, and Niloofar Ayoobi Yazdi[b]

[a]Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA 24061 USA

[b]Department of Radiology, University of Medical Sciences, Tehran 1419733141, Iran

ABSTRACT

The global extent of COVID-19 mutations and the consequent depletion of hospital resources highlighted the necessity of effective computer-assisted medical diagnosis. COVID-19 detection mediated by deep learning models can help diagnose this highly contagious disease and lower infectivity and mortality rates. Computed tomography (CT) is the preferred imaging modality for building automatic COVID-19 screening and diagnosis models. It is well-known that the training set size significantly impacts the performance and generalization of deep learning models. However, accessing a large dataset of CT scan images from an emerging disease like COVID-19 is challenging. Therefore, data efficiency becomes a significant factor in choosing a learning model. To this end, we present a multi-task learning approach, namely, a mask-guided attention (MGA) classifier, to improve the generalization and data efficiency of COVID-19 classification on lung CT scan images. The novelty of this method is compensating for the scarcity of data by employing more supervision with lesion masks, increasing the sensitivity of the model to COVID-19 manifestations, and helping both generalization and classification performance. Our proposed model achieves better overall performance than the single-task baseline and state-of-the-art models, as measured by various popular metrics. In our experiment with different percentages of data from our curated dataset, the classification performance gain from this multi-task learning approach is more significant for the smaller training sizes. Furthermore, experimental results demonstrate that our method enhances the focus on the lesions, as witnessed by both attention and attribution maps, resulting in a more interpretable model.

KEYWORDS

Convolutional Neural Network; COVID-19 Diagnosis; Mask-guided Attention; Multi-task Learning; Model Interpretability; Lung CT Scan Dataset.

*Corresponding author: Zhenyu (James) Kong. Email: zkong@vt.edu

1. Introduction

The coronavirus pandemic has struck the world since late 2019, causing a global crisis and countless deaths. As of January 11th, 2021, the World Health Organization has reported more than 308.46 million confirmed cases and 5.49 million deaths due to this virus. Furthermore, COVID-19 mutations have significantly constrained hospitals' capacity resulting in delayed care and increased risks for patients suffering from other critical conditions. COVID-19's global reach has brought together experts from a wide range of fields to combat the disease. One of the ongoing research topics has been to improve the COVID-19 diagnosis. Early diagnosis has two main benefits: (1) lowering the infectivity rate by isolating patients; and (2) reducing the fatality rate through early intervention.

While the reverse transcription-polymerase chain reaction (RT-PCR) test, which falls under the category of nucleic acid amplification tests (NAATs), has become the gold standard for detecting COVID-19, it has drawbacks such as limited sensitivity to the new variants, short supply of testing kits, and lengthy wait time for results Ai et al. (2020); Tahan et al. (2021); Trivizakis et al. (2020); Xie et al. (2020). Alternatively, lung computed tomography (CT) has proven to be a rapid and relatively accurate method of detecting COVID-19 and severity assessment Ai et al. (2020); Fang et al. (2020); Trivizakis et al. (2020); Xie et al. (2020). Infected patients' lung CT scans may exhibit distinctive characteristics such as ground-glass opacification, bilateral involvement, and diffuse distributions Misztal et al. (2020); Trivizakis et al. (2020); Xie et al. (2020). However, interpreting CT scans is a complex task requiring extensive radiology expertise. The number of radiologist experts is limited, and they face a heavy workload during an outbreak, increasing the risk of human errors. Therefore, transferring expert knowledge into intelligent models is valuable in order to improve healthcare accessibility, reduce the medical specialists' workload and their unnecessary exposure to the outbreak.

Deep learning has become one of the most extensively used approaches for building intelligent models, which can learn the underlying representation of images and classify them in a time-efficient manner. Notably, deep learning approaches have been successful for COVID-19 diagnosis in lung CT scans. Zhang et al. (2020) proposed an AI system that can identify COVID-19 markers and lesion properties using an extensive CT database of 3,777 patients. Zhao et al. (2020) used a mix of CT scans, lung, and lesion masks to train a COVID-19 diagnosis model leveraging multi-task and self-supervised learning. Rahimzadeh et al. (2021) presented a fast, accurate, and fully automated method for COVID-19 diagnosis from the patient's chest CT scan images. There have been several other studies on deep learning-based COVID-19 diagnosis

Maftouni et al. (2021); Polsinelli et al. (2020); Shamsi et al. (2021); Yazdani et al. (2020). Most of the work uses a single-task approach and devotes the learning model to only one task. On the other hand, jointly learning multiple related tasks, namely, multi-task learning (MTL), has been shown to overcome over-fitting and improve generalization by implicit data augmentation, attention focusing, and regularization Ruder (2017).

Despite the promising learning ability of deep models, the generalization power of the trained network depends on the size, distribution, and quality of the training dataset. Inadequate training datasets can easily lead to over-fitted deep learning models that cannot generalize well on a new dataset. Some COVID-19 datasets have been made publicly available Afshar et al. (2021); Cohen et al. (2020); He et al. (2020); Jun et al. (2020); MedSeg (2020); Morozov et al. (2020); Rahimzadeh et al. (2021); Zhao et al. (2020). Zhao et al. (2020) introduced the COVID-CT dataset, which includes 349 COVID-19 CT images from 216 patients and 463 non-COVID-19 (a mix of normal cases and patients with other diseases). Misztal et al. (2020) reported improving classification performance by categorizing negative COVID-19 cases into specific groups and creating the COVID-19 CT Radiograph Image Data Stock dataset with careful data split. Afshar et al. (2021) built an open-sourced dataset named COVID-CT-MD, comprising COVID-19, Normal, and community-acquired pneumonia (CAP) cases. The COVID-CT-MD is accompanied by lobe-level, slice-level, and patient-level labels to aid in developing deep learning methods. Notwithstanding, researchers continue to require more data for deep learning models' training in order to provide better insights and generalization performance. To this end, our COVID-19 lung CT-scan dataset is curated from seven open-source datasets.

Our proposed method applies a deep learning model with an attention module, which is the state-of-the-art technique in machine learning, to improve the performance of COVID-19 detection. For an image input, the attention module infers the attention map, which is a collection of pixel-level weights, to prioritize the image features by the level of importance for the task Woo et al. (2018). It attempts to mimic human visual perception that focuses on specific locations, objects, and attributes in the scene by filtering out irrelevant information. For example, an expert radiologist knows precisely where to focus in a CT scan to find a particular pathology. So, intuitively, the attention map learns which areas on the image are more relevant to the performed task, such as medical diagnosis. The use of attention modules in deep learning networks originated and proved successful in neural machine translation Bahdanau et al. (2014); Vaswani et al. (2017). Motivated by this success and its consistency with human perception, visual attention modules were adopted in different computer vision applications such as image

captioning Xu et al. (2015), visual question answering Lu et al. (2016), and image classification Wang et al. (2017). The Residual Attention Network in Wang et al. (2017) achieved state-of-the-art object recognition performance on several benchmark datasets and showed improved robustness against noisy labels. Later, Woo et al. Woo et al. (2018) proposed a lightweight convolutional block attention module (CBAM) that could be integrated into any convolutional neural network (CNN) architecture to infer and refine attention. They showed that integrating CBAM inside various state-of-the-art CNN models improves the classification and detection performance. Accordingly, CBAM is incorporated into our model for enhanced performance through attention map learning and feature refinement.

To summarize, the objective of this paper is to improve the generalization and performance of COVID-19 detection deep learning models. Specifically, the main contributions of our paper are as follows:

- A large and broadly representative lung CT scan dataset for COVID-19 detection is built by curating seven open-source datasets. To the best of our knowledge, this is the largest publicly available COVID-19 CT dataset, accompanied by patient metadata. The dataset includes cases from 13 countries and has three classes: COVID-19, Normal, and CAP. The dataset also consists of COVID-19 frames with corresponding lesion masks merged from three of the datasets.
- A novel mask-guided attention (MGA) classifier for COVID-19 diagnosis is developed that improves classification performance, data efficiency, and interpretability. Our experimental results demonstrate the proposed method's superior performance over the baseline and improved focus on the COVID-19 lesions.

The remainder of this paper is organized as follows. In Section 2, a brief review of related research work on COVID-19 diagnosis, lesion segmentation, MGA methods, and multi-task learning is provided. Next, the proposed research methodology is summarized in Section 3. Section 4 introduces our curated CT scan dataset. Our proposed MGA deep learning model for COVID-19 diagnosis is detailed in Section 5, followed by the experimental results and ablation studies in Section 6. Finally, the conclusions and future directions are discussed in Section 7.

2. Related Work

The related works in deep learning-based COVID-19 diagnosis and lesion segmentation on CT scans is reviewed first in Section 2.1. Next, the multi-task learning related to COVID-19 are introduced in Section 2.2. The research gap is identified in Section 2.3.

2.1. COVID-19 Diagnosis and Lesion Segmentation based on CT Scans

Deep learning has been the method of choice in most existing works on diagnosing COVID-19 infection from CT scans He et al. (2020); Polsinelli et al. (2020); Rahimzadeh et al. (2021); Shamsi et al. (2021); Yazdani et al. (2020); owing to the success of deep learning methods in image classification. He et al. (2020) tested seven state-of-the-art deep classification models including VGG-16 Simonyan and Zisserman (2014), ResNet18, ResNet-50 He et al. (2016), DenseNet-121, DenseNet-169 Huang et al. (2017), EfficientNet-b0, and EfficientNet-b1 Tan and Le (2019). They integrated contrastive self-supervision Chen et al. (2020) into the transfer learning process to further improve the performance of deep classification algorithms. In Rahimzadeh et al. (2021), a two-stage system was proposed for detecting COVID-19. The first stage filtered out those CT frames in which the inside of the lung is not properly observable. At the next stage, they applied a new feature pyramid network designed for classification problems using a ResNet-50V2 baseline He et al. (2016), allowing the model to investigate different resolutions of the image and maintain the data from small objects. Polsinelli et al. (2020) proposed a light Convolutional Neural Network design, based on the SqueezeNet model Iandola et al. (2016), for the efficient differential diagnosis of COVID-19 CT scans from other community-acquired pneumonia infections and healthy CT scans. Shamsi et al. (2021) proposed a novel transfer learning-based and uncertainty-aware framework for reliable detection of COVID-19 cases from X-ray and CT images. In Yazdani et al. (2020), the attentional convolution network Wang et al. (2017) is proposed to focus on the infected areas of the chest so that the network can provide a more accurate prediction.

Lesion segmentation is another task on CT scan images that is well suited for deep learning Chaganti et al. (2020); Chassagnon et al. (2020); Gao et al. (2021); Wu et al. (2021); Yao et al. (2021). Generally, this task entails automatically predicting binary lesion masks, assigning the same label to all types of lesions. The problem can be expanded to the semantic segmentation of different types of lesions and within and outside lung regions if a sufficient number of lesion-specific ground truth masks are available. Nonetheless, the binary lesion masks are adequate for assessing the extent of involvement and manifestations of the disease in the lung of a confirmed or suspected COVID-19 patient Tilborghs et al. (2020). Chaganti et al. (2020) proposed to automatically segment ground-glass opacities (GGO) and areas of consolidation together using a DenseUNet Ronneberger et al. (2015). Chassagnon et al. (2020) proposed CovidENet: an ensemble of 2D and 3D CNNs based on AtlasNet Vakalopoulou et al. (2018) for binary lesion segmentation and achieved human-level segmentation performance in terms of Dice Score and

Hausdorff distance. Yao et al. (2021) proposed the NormNet, a voxel-level anomaly modeling network to recognize normal voxels from possible anomalies. A decision boundary for normal contexts of the NormNet was learned by separating healthy tissues from the diverse synthetic "lesions," which can segment COVID-19 lesions without training on any labeled data. To focus more on the lesion areas, a novel lesion attention module was developed to integrate the intermediate segmentation results.

2.2. Multi-task Learning (MLT)

In general, MTL is known as a machine learning approach that assimilates information from correlated tasks to improve the generalization capability of the overall learning model Zhang and Yang (2017). There are two approaches in multi-task learning: hard parameter sharing and soft parameter sharing of hidden layers Ruder (2017). The hard parameter sharing is commonly found in the literature, in which multiple tasks (networks) share some hidden layers while keeping their separated output layers. On the other hand, soft parameter sharing is achieved when each task has its separate model and respective parameters, but the parameters from different tasks are jointly regularized.

MTL has been adopted for COVID-19 diagnosis improvement. Bao et al. (2020) proposed end-to-end multi-task learning to detect and assess the severity of COVID-19 cases with improved performance using only a relatively small dataset of 1329 CT scans. Amyar et al. (2020) developed a multi-task deep learning model with three tasks of classification, segmentation, and reconstruction from chest CT images. Goncharov et al. (2021) deployed a two-task deep learning model to identify COVID-19 cases and quantify the disease severity. Wu et al. (2021) developed a novel Joint Classification and Segmentation (JCS) system to perform real-time and explainable COVID-19 chest CT diagnosis. Gao et al. (2021) developed a dual-branch combination network (DCN) for COVID-19 diagnosis to simultaneously achieve individual-level classification and lesion segmentation. These papers reported an improvement over the single-task benchmark models. Furthermore, multi-task learning has improved the performance of smaller datasets more significantly Crichton et al. (2017); Gong et al. (2019).

Another form of MTL is MGA models that extend the attention convolutional neural networks. The attention weights that the model assigns to each input element are generally learned without dedicated supervision; therefore, they might also converge to irreverent parts of the image for the task. For example, in classifying lung CT scans, the main focus should be on the inside lung manifestations, and assigning high attention weights to outside lung pixels is useless. Accordingly, recent research adopts extra supervision on attention map training. For

instance, Song et al. (2018) designed a contrastive attention model guided by binary masks. It can generate a pair of body-aware and background-aware attention maps, which can produce features of body and background for Person Re-Identification. Pang et al. (2019) introduced a novel MGA network that fits into popular pedestrian detection pipelines. The attention network emphasizes visible pedestrian regions while suppressing the occluded parts by modulating full body features. Wang et al. (2021) proposed an MGA model that provides auxiliary supervision from predicted masks from a pre-trained segmentation model for discriminative and patchy representation learning.

2.3. Research Gap

The research gaps in the COVID-19 diagnosis approaches listed in Sec. 2.1 and Sec. 2.2 are identified as: (1) most proposed COVID-19 diagnosis methods are single-task, which may be more susceptible to over-fitting; (2) Training an accurate COVID-19 diagnosis model requires a large amount of broadly representative sample data, which many of the existing research efforts lack; and (3) Some COVID-19 diagnosis applications using a multi-task approach demonstrated improved performance over the single-task model; however, their models lack explainable choices of diagnosis results. In summary, there is a need for a novel approach to improve the generalization, interpretability, and data efficiency of the deep learning model for COVID-19 diagnosis applications. Therefore, in this work, a multi-task COVID-19 detection model, jointly supervising the attention maps and class labels, is developed to fill the aforementioned research gaps. The multi-task learning approach is implemented through an MGA module integrated inside the COVID-19 classifier to supervise its attention map with segmented lesions. Additionally, one of the strengths of our model is that it is trained and tested on a more broadly representative dataset, which promotes its generalizability.

3. Proposed Research Methodology

This work aims to develop a data-efficient deep learning model for COVID-19 diagnosis based on chest CT scan slices, with good generalization and interpretability. The performance of the deep learning model is highly dependent on the training data. However, a comprehensive CT scans dataset for COVID-19 is not publicly available to the researchers in the current literature. To fill this gap, a new CT scans dataset for COVID-19 is created and introduced in Section 4.

A CT scan cross-section or slice is reconstructed from the measurements of attenuation coefficients (intensity reduction) of x-ray beams as it passes through the tissues. Tissues with higher attenuation (such as bones) are bright, whereas tissues with little attenuation (such as

air and water) appear dark. Since a normal lung looks dark in the CT scan, the abnormal increase in the attenuation in an inside lung area points at lesions related to different diseases (e.g. COVID-19 or CAP). Radiologists have characterized the key lung lesion patterns, or lesion types, for COVID-19 diagnosis. Our dataset contains the marking of these patterns as follows,

(1) ground-glass opacities (hazy gray opacities that do not obscure the underlying vessels),
(2) consolidation (areas of increased attenuation that obscure the underlying vessels), and
(3) pleural effusion (excess fluid build-up between the lung and chest cavity).

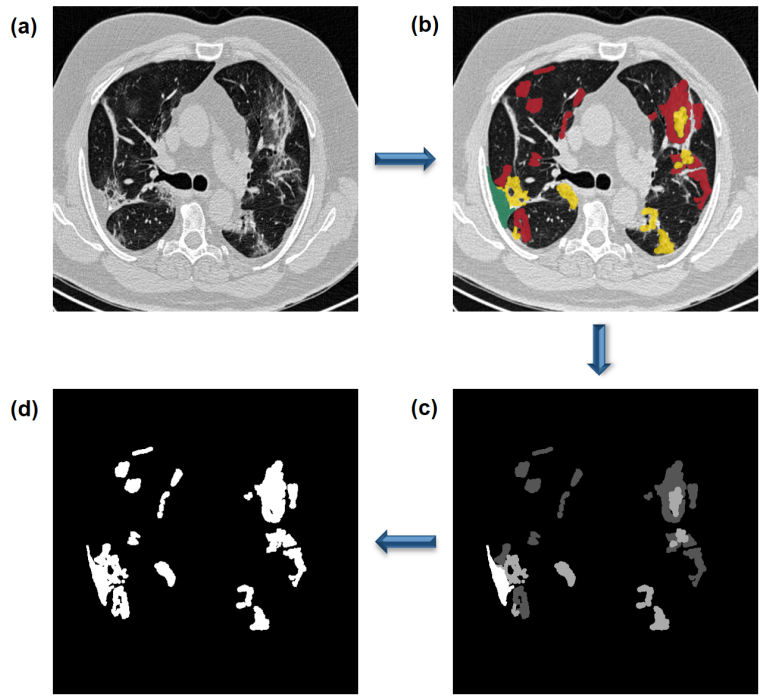


Figure 1. Example of Deriving binary lesion mask from radiologist annotations on a chest CT scan slice. (a) The chest CT scan slice (b) The radiologist annotation from MedSeg (2020). Red, yellow, and green colors indicate ground-glass opacities, consolidation, and pleural effusion lesion types, respectively. (c) Semantic segmentation mask that maps non-lesion pixels to black and assigns each lesion type to a different class (level of gray) (d) Binary (black and white) lesion mask, after mapping all lesion types to the general category of lesions. Black represents non-lesion, and white represents lesion.

These patterns are revealed through the lesion annotations manually marked by radiologist experts. As depicted in Figure 1, the lesion annotations are employed to derive the binary lesion masks of each image. Namely, black pixels are non-lesion while white ones are lesions. Therefore, in this paper, all different COVID-19 lesion types are combined as one type used in the classification analysis for COVID-19 diagnosis.

Our idea is to fully utilize the available domain knowledge through the COVID-19 lesion patterns to improve the deep learning model performance while lowering its data requirement. The overall proposed model architecture, depicted in Figure 2, is a two-step approach as follows.

- Step 1 (Section 5.1): A lesion segmentation model based on Hierarchical Multi-scale At-

tention Network Tao et al. (2020) (HMSANet) is implemented to automatically create lesion masks for the images that radiologists did not mark with lesion masks since manual marking is costly and time-consuming. The lesion masks are then used in the MGA module to supervise the spatial attention map (created by CBAM in Step 2) for the purpose of assigning higher attention weights to the pixels resembling lesions.

- Step 2 (Section 5.2): The deep learning classification model is applied to classify the input CT image, guided with the lesion mask generated in Step 1, and provides the diagnosis result, namely, Normal, COVID-19, or CAP case. Our classification model uses CBAM and MGA modules to enhance the model's focus on lesion locations. Particularly, the spatial attention map created by CBAM is guided towards the lesions through the MGA module during training.

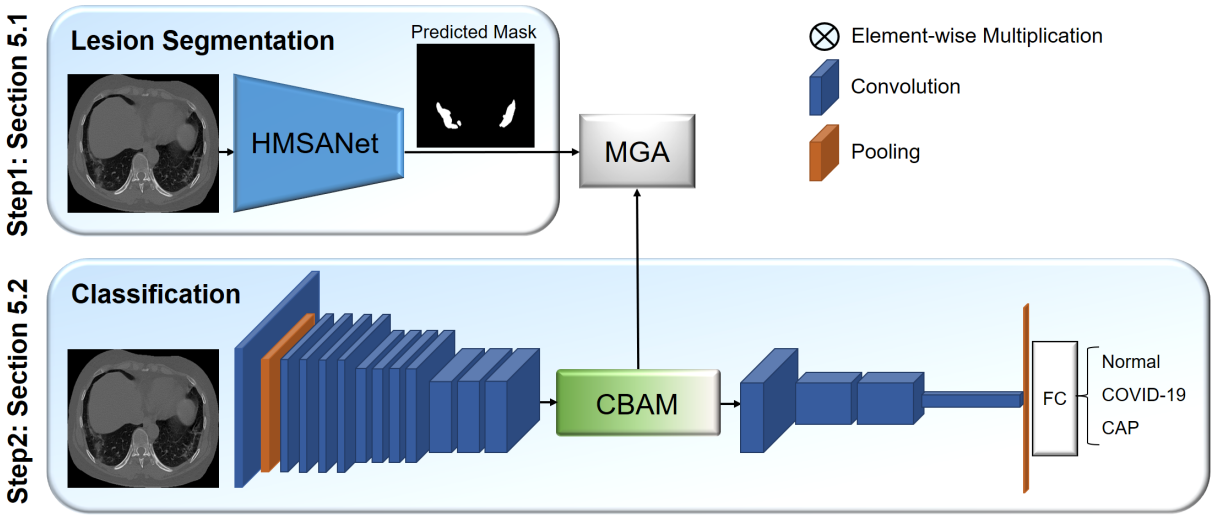


Figure 2. The proposed method architecture with two main parts: Lesion segmentation and classification

These two steps introduced above are integrated through the hard parameter sharing multi-task learning model (namely, the share of some hidden layers). The first task, accomplished through the MGA module, directly supervises the network's attention map using the lesion masks predicted in Step 1. The second task, implemented in the second step, applies supervision on the class predictions with the ground-truth class labels. This multi-task learning model has the following advantages.

(1) First, the increased focus on the lesion regions, which are the COVID-19 manifestations, improves the accuracy of COVID-19 diagnosis and alleviates over-fitting by lowering the effective dimensionality of the data.

(2) Second, it lowers the training data requirement. Our experiments show that the proposed

model offers fewer training data sample requirements by utilizing additional supervision through the lesion data.

(3) Third, the model prediction is more interpretable and reliable when focusing on the lesions instead of the entire image with many irrelevant parts to the illness.

4. Dataset Creation

CT scans show promise in providing COVID-19 screening and testing accurately and quickly Zhao et al. (2020). We created a large lung CT scan dataset for COVID-19 to aid in developing the diagnosis models. The dataset includes curated data from Afshar et al. (2021); Cohen et al. (2020); Jun et al. (2020); MedSeg (2020); Morozov et al. (2020); Rahimzadeh et al. (2021); Zhao et al. (2020). Each of the seven datasets is illustrated by an example image in Figure 3. These datasets have been utilized publicly in COVID-19 diagnosis literature and have proven effective in deep learning applications. As a result, the combined dataset is expected to increase the generalization capacity of deep learning models by learning from all of these resources together.

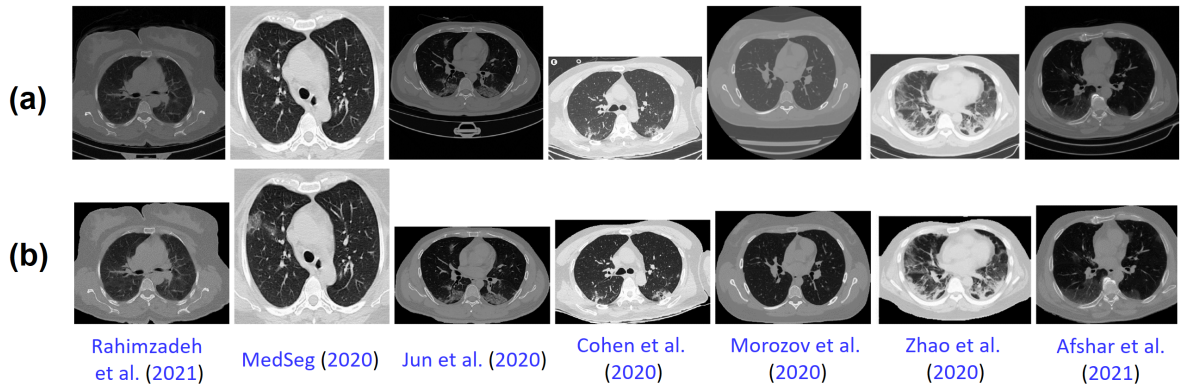


Figure 3. (a) An example lung CT frame from the seven open-source datasets included in our dataset; (b) Same images after initial preprocessing, including background removal, cropping, and normalization on the foreground segment.

Our objective is to provide a large dataset of axial chest CT scan slices with three labels, namely, (1) COVID-19, (2) Normal, and (3) CAP, together with their corresponding metadata and lesion masks if available. The CT scan cross-sections or slices (also referred to as frames or images) can be layered on top of each other to create a 3D volume. While some of these seven public datasets consist of class labeled CT slices, Others include 3D CT scan volumes with various slice-level annotations. We used these annotations to extract and label the CT slices from the CT volumes. Additionally, these datasets have different data formats, such as NIfTI, DICOM, TIFF, PNG, and JPG. All the extracted CT scan slices are converted to 8-bit PNG file format for better uniformity and accessibility for the deep learning analysis (See Figure 3).

It should be noted that not all of the 3D CT volumes in the dataset were annotated with class

Table 1. Seven Datasets summary.

Dataset	Country	COVID-19 Slices	COVID-19 Cases	Normal Slices	Normal Cases	Masks	Gender & Age	Data Format
Zhao et al. (2020)	China, Japan	349	213	NA	NA	NA	Missing	PNG, JPG
Afshar et al. (2021)	Iran	3,815	55	760	76	lobe level annotation	Available	DICOM
Cohen et al. (2020)	Multiple	34	17	NA	NA	NA	Missing	PNG, JPG
Morozov et al. (2020)	Russia	785	50	5,080	254	lesions	NA	NIfTI
Rahimzadeh et al. (2021)	Iran	666	68	1,053	274	NA	Available	TIFF
Jun et al. (2020)	Multiple	1,844	20	NA	NA	lung and lesions	Available	NIfTI
MedSeg (2020)	Italy	100	43	NA	NA	lung and lesions	Available	NIfTI
Ours	**Multiple**	**7,593**	**466**	**6,893**	**604**	**64% Missing**	**9% Missing**	**PNG**

Note: NA stands for not available, and Missing is available but with missing values.

labels at the slice level, and we worked with our radiologist to annotate the remaining CT images. To ensure the dataset quality, we excluded the chest slices that do not carry information about inside lung manifestations, as well as the adjacent slices with almost identical appearances. Additionally, we removed images lacking clear class labels or patient information. We have collected 7,593 COVID-19 images from 466 patients, 6,893 normal images from 604 patients, and 2,618 CAP images from 60 patients in total. Our CAP images are all from the dataset Afshar et al. (2021), in which 25 cases are already annotated. Our radiologist has annotated the remaining 35 CT scan volumes.

Table 1 summarizes the number of frames from COVID-19 and normal classes, the availability of specific metadata and masks, and the initial data format of each of the seven datasets. As previously stated, all of the cases have patient ID, necessary for data splitting. As listed in the table, three of the datasets have lesion masks Jun et al. (2020); MedSeg (2020); Morozov et al. (2020), providing us with 2,729 COVID-19 lesion masks (36% of the COVID cases) to be used to train the mask segmentation model, explained in Section 5.1. The distinct categories of lesions in MedSeg (2020) are mapped to a binary lesion mask for consistency across datasets.

Figure 4 depicts multiple statistics from the dataset. The country and gender distributions on the entire dataset are shown in the subfigures (a-b). Figure 4(a) indicates that the cases come from 13 countries, with Iran, Russia, and China ranking first through third. According to Figure 4(b) most of the cases are male, and this male dominance holds for all Normal, COVID-19, and Cap classes. Figure 4(c) compares the age distribution of the three classes and shows that all the age groups are represented in the dataset. The median age of Normal, COVID-19, and CAP classes are 50, 49, and 59, respectively. Figure 4(d) compares the prevalence of distinctive CT characteristics in the 796 COVID-19 cases with CT scan reports, highlighting that ground-glass opacities, bilateral involvements, and consolidation have frequently been reported. And patterns attributed to higher severity, such as diffuse distribution Lei et al. (2021), are also present. These statistics indicate that the dataset population is broad and representative, having cases from various ages, gender, nationality, and severity groups.

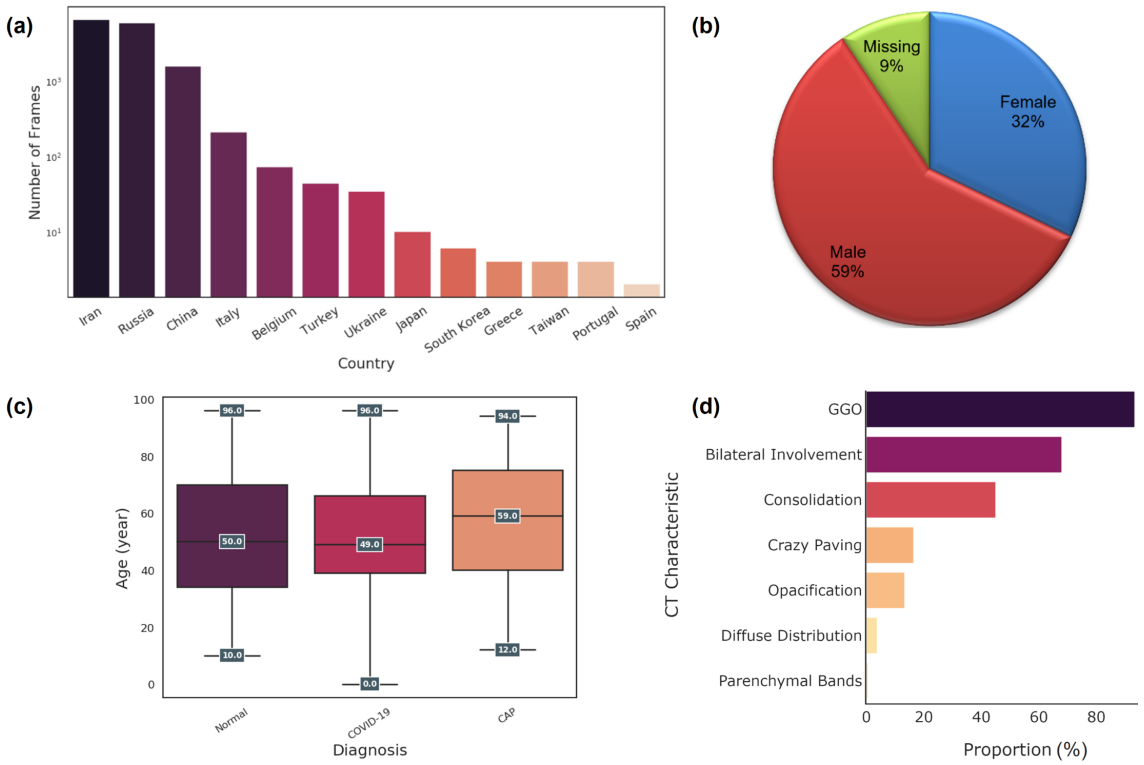


Figure 4. (a) Country and (b) Gender distribution of images in our dataset (c) Comparison of the three classes' age ranges (d) The proportion of critical COVID-19 manifestations in the available CT scan reports.

5. COVID-19 Diagnosis Using Deep Learning with MGA Model

This section presents the multi-task learning model using MGA in detail, consisting of two steps (See Figure 2). Step 1: in Section 5.1, a lesion mask prediction model is implemented based on the 2729 COVID-19 available lesion masks and then applied to generate the lesion mask in all the images that were not annotated with lesions. Step 2: in Section 5.2, a classification model is developed to classify if the input image is Normal, COVID-19 or CAP. Additionally, the significance and method of interpreting the model predictions is introduced in Section 5.3.

5.1. Segmentation Model for Lesion Mask Prediction

Semantic segmentation is to classify every pixel in the image into one of the classes of interest. The problem in this paper is simplified to binary segmentation when the aim is to separate out a single class, namely, lesions. Segmentation may be thought of as a pixel-wise classification that requires object localization and boundary detection at the same time.

Localization and boundary detection require different image resolutions and network receptive fields (the extent of an image exposed to a single neuron within the model). Predicting object location is better handled at a scale-down image size because the network's receptive field can observe more of the image context. In contrast, detecting fine edges and thin structures is better handled at a scaled-up image size, leading to a smaller receptive field. Therefore,

multi-scale inference is an effective means to address both of these underpinning segmentation requirements. The challenge is how to combine the multiple-scale predictions effectively. The simplest way is to combine the results with averaging or max pooling. A more effective approach is to find the weighted average of the multiple scale-level predictions based on pixel-level weight maps learned within the model. HMSANet Tao et al. (2020) uses the second approach and hierarchically combines the multiple scale predictions using the learned weight map, also called attention map. This model can learn the relative weighting between adjacent scales during training and enables the inclusion of other scales during inference on the test images.

HMSANet is adopted in this study for the lesion segmentation because its multi-scale and high resolution learning facilitates lesion localization and accurate boundary detection, especially as lesions appear in different sizes and shapes. Additionally, our results presented in 6.1 shows that HMSANet outperforms other segmentation methods on the lesion segmentation task.

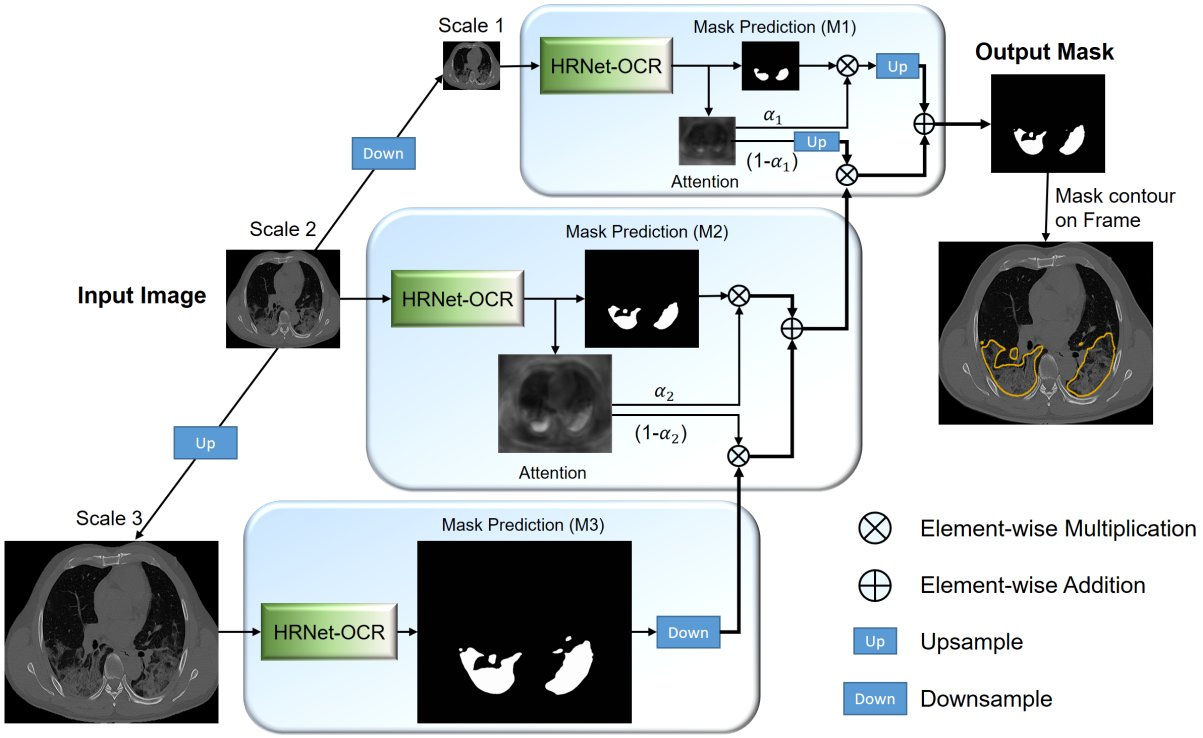


Figure 5. The lesion segmentation model (the adopted HMSANet module Tao et al. (2020) structured in Figure 2). HMSANet infers the lesion mask of the same size as the input image by hierarchically combining predictions at multiple scales, weighted by the hierarchically learned attention weights. Lower scales determine the general lesion location, while higher scales refine its details and edges.

The way that the HMSANet model is adopted for lesion segmentation is shown in Figure 2 (upper part), which is the first step of the proposed methodology. HMSANet model structure is depicted in Figure 5 in which the lesion mask is inferred using three frame scales. These image scales pass through a network trunk for both scale-level lesion mask and attention map infer-

ence. The High-Resolution network Object-Contextual Representations (HRNet-OCR) model Yuan et al. (2020) is the best-performing scale-level trunk for the HMSANet model showing competitive performance on several semantic segmentation benchmarks. As shown in Figure 5, these scale-level mask predictions are combined to generate the final lesion mask by applying a chain of element-wise multiplication between the attention maps (α_n) and the mask predictions (M_n), followed by element-wise addition among the multiple scales. The chain starts at the lowest scale of the image, namely scale 1 in Figure 5, which captures the most global features, and is further refined for details at the following higher scales in order (Scale 2 and 3). Since lower scales take precedence, they take out their contribution share ($0 < \alpha_n(i,j) < 1$), higher (whiter) at the pixels of increased confidence, and pass the remaining attention ($1 - \alpha_n(i,j)$) to the following higher scales. Specifically, the final predicted mask ($\mathbf{M}$) is calculated by Equation (1), in which $\mathbf{U}$ is bilinear upsampling and $\mathbf{D}$ is downsampling.

$$\mathbf{M} = \mathbf{U}(\alpha_1 \otimes \mathbf{M}_1) + \mathbf{U}(1 - \alpha_1) \otimes [(\alpha_2 \otimes \mathbf{M}_2) + ((1 - \alpha_2) \otimes \mathbf{D}(\mathbf{M}_3))] \tag{1}$$

The HMSANet model is trained using the cross-entropy loss function, batch size of 1 per GPU, image scales of 0.5 (scaled down to half the size) and 1.0, stochastic gradient descent optimizer with the learning rate of 0.01, the momentum of 0.9, and the weight decay of $5e^{-4}$. These segmentation model hyperparameters are determined based on their values in the base paper Tao et al. (2020), achieving a new state-of-the-art performance, and showed the best performance on our validation set. We used four NVIDIA GeForce RTX 2080 Ti GPUs and Pytorch library to train the model. The 2729 COVID-19 frames and their ground truth lesion masks are split into the training, validation, and test sets in sizes of 2329, 200, and 200, respectively.

After evaluating the segmentation performance, the trained segmentation model is employed to predict COVID-19 lesions masks on all the images without lesion masks, regardless of their class. Then, all the images are paired with their corresponding masks to be used as the ground-truth of the network's attention map in the MGA module, as laid out in the next section.

5.2. Classification Model for COVID-19 Diagnosis

The lightweight Residual Network He et al. (2016) with 18 layers (ResNet18) is selected in this work to serve as the backbone of our COVID-19 classification architecture. The residual networks resolve the vanishing gradient and performance degradation problems of deep networks through skip connections, also known as residual connections. Specifically, Resnet18 is chosen for its lightweight architecture, computational efficiency, and competitive performance in COVID-19 diagnosis Helwan et al. (2021); Pham (2020). The ResNet18 architecture is our baseline model

but without attention. We have embedded CBAM Woo et al. (2018) as the attention module in the ResNet18 architecture to enhance the activation of discriminate parts of the input image. For the second step of the proposed methodology, namely, the lower part of Figure 2, the more detailed structure is shown in Figure 6.

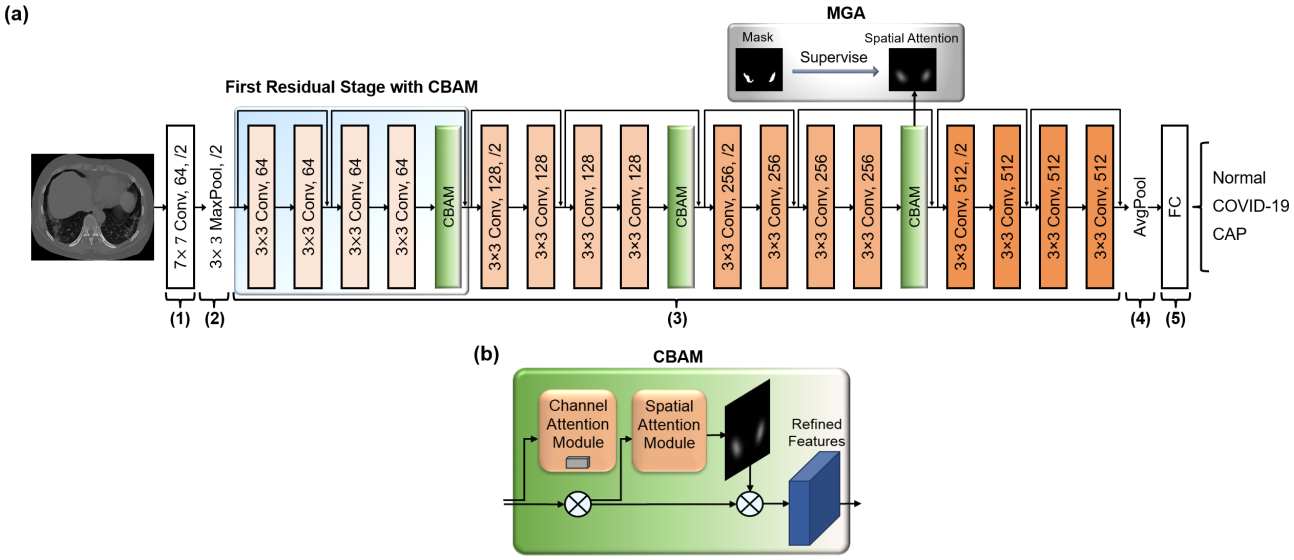


Figure 6. (a) The classification model's network structure (b) CBAM structure

our classification model's network structure consists of the following components:

(1) a convolutional layer (with 7×7 filter size, 64 filters, and stride of 2) to learn 64 filters,

(2) a max pooling layer (with 3×3 filter size, and stride of 2) to reduce the input spatial size,

(3) four residual stages (four successive convolutional layers with two residual connections and the same number of filters, distinguished by the color in Figure 6), to allow the information flow between layers while gradually reducing the spatial size and learning more filters. CBAM is embedded only in the first three residual stages to save the computation,

(4) an average pooling layer, to spatially down-sample the feature map into a vector, and

(5) a fully connected layer at the end for classification.

Each convolutional layer outputs a 3D tensor called a feature map with (height, width) as the spatial axes and multiple-output channels (C) based on the number of filters. The feature maps of convolutional layers in each residual stage have the same dimension. From one residual stage to the next, the feature maps' height and width are halved (noted by /2 in Figure 6) by convolution stride, and the output channels are doubled (64, 128, 256, and 512, respectively). The attention module's role is to reweight the feature map. Since the feature maps are 3D tensors, the feature map re-weighting can be performed spatially (by spatial attention module) or on the channels (by channel attention module). The spatial attention module assigns higher

weights to more informative parts of the input, while the channel attention module weights the channels based on their relevance and importance by multiplying the channel weights with the feature map. CBAM has a consecutive channel and spatial attention (sub)modules, which is shown to be the best performing combination.

The ResNet18 model with embedded CBAM is our baseline with attention but without direct supervision of attention map learning. In addition to applying attention reweighting, our proposed multi-task model uses an MGA module to directly supervise the spatial attention map of one of the three CBAMs by the predicted masks. Figure 6 shows our classification model's network structure when the MGA module is placed at the third residual stage. The optimal placement of the MGA module has been studied in Section 6.3.

In order to create the spatial attention map, the spatial attention module average-pools and maximum-pools the channel-attended feature map of dimension (H,W,C) to aggregate and squeeze its channel information into two (H,W,1)-dimensional tensors. Then, these two poolings are concatenated in the channel dimension (H,W,2), and transformed into the spatial attention map via a convolutional layer with one channel output, padding of 3, filter size of 7×7 ($f^{7\times7}$), and a sigmoid activation function (σ), as formulated in Equation (2). Therefore, the spatial attention map is a one-channel tensor with the same height and width as its corresponding feature map (H,W,1) in which all the values are between zero and one.

$$SA(f_i) = \sigma(f^{7\times7}([AvgPool(f_i); MaxPool(f_i)])) \tag{2}$$

As indicated by Equation (3), the image features extracted at the j^{th} residual stage denoted by $f_j \subseteq \mathbb{R}^{H\times W\times C}$ are spatially multiplied by the spatial attention map $SA_1 \subseteq \mathbb{R}^{H\times W}$ to construct the attended features ${f_j}^{att}$. H, W, and C denote height, weight, and the number of channels, respectively. In the element-wise multiplication of the broadcasted (copied) one-channel spatial attention map with multi-channel features, i signifies the channel index.

$${f_j}^{att}(i) = f_j(i) \otimes SA(f_{j(i)}) \tag{3}$$

We directly supervise one of the spatial attention maps (SA) with the same sized predicted lesion mask (M) from Step 1 (section 5.1) by minimizing the pixel-wise mean squared error loss function L_{att}:

$$L_{att} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \|M_{i,j} - SA_{i,j}\| \tag{4}$$

The MGA module is intended to direct the spatial attention map emphasis to the inside lung manifestations and give extra attention to lesions and lung parts that resemble lesions. Since

the predicted masks might not completely match the ground truth lesion masks, the residual connection facilitates the flow of information by adding the initial features via skip connections.

The classification task is supervised with the cross-entropy loss between the predicted class probabilities ($\hat{y}$) and one-hot encoded ground truth class labels (y) of the three classes as stated in Equation (5).

$$L_{ce}(\hat{y}, y) = -\sum_{k}^{3} y^{(k)} \log\, \hat{y}^{(k)} \tag{5}$$

As represented in Equation (6), we adopted multi-task learning with uncertainty loss weighting Kendall et al. (2018) between the classification cross-entropy and attention mean squared error losses because it has shown superior performance over using fixed weights Gong et al. (2019). This weighting scheme lets the model adjust the weight of each loss by learning the observation noise parameters σ_1 and σ_2 alongside the model weights (W). Smaller values of the observation noise parameter will increase the contribution of its associated loss function. These noise parameters are regularized to avoid very large values, which diminishes the contribution of each of the tasks.

$$L(W, \sigma_1, \sigma_2) = \frac{1}{2\sigma_1^2} L_{ce}(W) + \frac{1}{2\sigma_2^2} L_{att}(W) + \log\sigma_1 + \log\sigma_2 \tag{6}$$

The model is trained using an Adam optimizer with a learning rate of 0.0001, a cosine annealing scheduler, a batch size of 32, and 100 epochs with early stopping with the patience of 10. These hyperparameters are tuned using Bayesian Optimization Nogueira (2014). We used four NVIDIA GeForce RTX 2080 Ti GPUs and Pytorch library to train our models.

Table 2. Train, Validation, and Test splits distribution.

Data Split	COVID-19	Normal	CAP	Total
Train	5,563	4,643	1,773	11,979
Validation	1,508	1,736	643	3,887
Test	522	514	202	1,238

Table 2 specifies the dataset split between train, validation, and test. All the data splits are made in a patient-aware and stratified manner to avoid performance overestimation. Patient-aware splitting means keeping images from each unique patient together in one of the sets. Stratification implies that the splitting has the same proportion for each of the classes. Patient-aware splitting must be strictly adhered to. Limited by the patient-aware splitting, stratification is performed as much as feasible.

5.3. Interpreting the Model's Prediction

So far, we have introduced our proposed classification model that provides the COVID-19 diagnosis prediction but without interpretability. Achieving highly accurate but uninterpretable decisions makes deep learning models less trustable and has an adverse impact on their clinical applications. Although deep learning has a black box nature, much recent work has investigated the flow of information and input-output connections in deep neural networks to shed light on how it predicts. Such explanation methods help increase trust in the model when it predicts correctly and identifies the failure modes (such as data corruption and learning wrong patterns) when wrong. The gradient-based attribution methods Shrikumar et al. (2017); Simonyan et al. (2013); Sundararajan et al. (2017); Zeiler and Fergus (2014) provide input-specific explanations of the deep learning predictions by assigning an attribution value to each input feature. Each gradient-based attribution method has a slightly different formulation for identifying the contribution of each feature to the model's output through backpropagating the output prediction and decomposing it on the input image. The result is an attribution map, an image with the same size as the input containing the pixel level contribution scores.

Attribution maps are often shown as heatmaps, representing the attribution map with colors. For instance, red indicates features that contribute positively to the activation of the target output; blue color distinguishes features that have a suppressing effect on it; and the white color indicates the insignificance for the derived output. In this work, we use two prominent attribution methods called Integrated Gradient Sundararajan et al. (2017) and DeepLIFT Shrikumar et al. (2017) methods to highlight disease features in the CT images. The Integrated Gradient method calculates the integral of gradients of each feature along the path from a baseline (such as a black image) to input, while DeepLIFT is its faster approximation.

6. Results and Discussion

This section presents the performance of the lesion mask segmentation method (Section 6.1) and the proposed classification model with attention (Section 6.2). Additionally, Section 6.3 covers the ablation studies to determine the placement of the MGA module and the effectiveness of MGA classification with different training set sizes. Finally, Section 6.4 presents the interpretability of the decisions of our deep learning model.

6.1. Segmentation Performance

The HMSANet architecture, presented in Section 5.1, is employed as the mask prediction method because of its state-of-the-art segmentation performance. We compared the HMSANet's

performance with UNet Ronneberger et al. (2015), SegNet Badrinarayanan et al. (2017), and DeepLabV3 Chen et al. (2017) architectures, which are among the most widely used segmentation methods in the literature. As reported in Table 3, HMSANet achieves the highest intersection over union (IOU), Dice coefficient, precision, and recall on the test set (consisting of 200 COVID-19 frames). Figure 7 provides the qualitative comparison of the predicted masks on five sample test images. According to this figure, HMSANet predicted masks most closely resemble the ground truth masks and are our best choice for the lesion prediction.

Table 3. Lesion segmentation performance Comparison (the best performance in each metric is in red).

Method	IOU (%)	Dice (%)	Precision (%)	Recall (%)
SegNet	43.48	60.60	50.00	76.92
UNet	54.05	70.17	68.50	71.94
DeepLabV3	70.42	82.64	81.96	83.33
HMSANet (adopted in this paper)	**74.63**	**85.47**	**84.03**	**86.96**

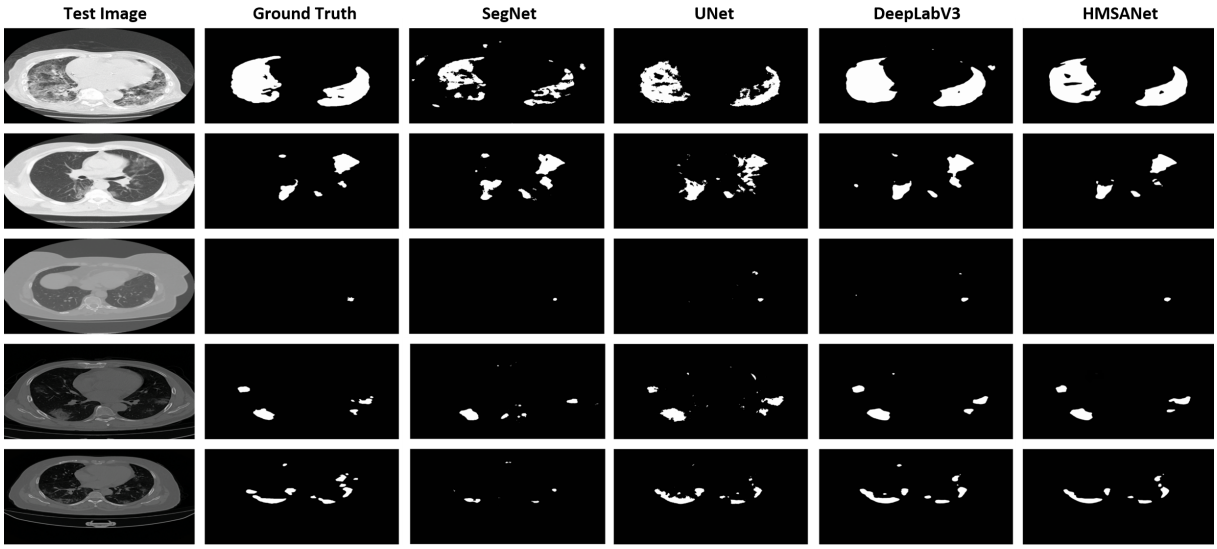


Figure 7. Lesion mask prediction comparison on five test images. HMSANet most closely resembles the ground truth.

6.2. Classification Performance

Table 4 compares our proposed classification model (Section 5.2) with two baseline models without and with attention modules (ResNet18 and ResNet18 + CBAM), and two state-of-the-art models (ResNet50 and DenseNet121). For the sake of consistency, all the models are trained from scratch. It is clear from the results that the overall performance of our proposed model is better than the other tested models. Particularly, the recall, also known as sensitivity, has shown the most significant improvement since the better focus on the lesions has boosted the detection of COVID-19 cases. In terms of recall, our model is the best and outperforms the second and third best-performing methods by 1.59% and 3.31%, respectively.

The ROC curve measures the true-positive rate (sensitivity) and false-positive rate (1 – specificity) trade-off, and its area under the curve (ROC AUC, also referred to as AUC) has meaningful interpretation for disease classification and is extensively used in medical diagnosis. F1 score, which is the harmonic mean of recall and precision, is another reported metric. Our model's enhanced AUC and F1 score metrics indicate that the increased sensitivity did not come at the expense of more false positives. The proposed multi-task learning improves generalization by leveraging the domain-specific knowledge contained in the training data and makes it capable of learning a more meaningful representation.

Table 4. Classification performance results (the best performance in each metric is in red).

Method	Accuracy (%)	F1 score (%)	Recall (%)	Precision (%)	ROC AUC (%)	time (s)
ResNet18	93.05	91.22	85.63	**97.60**	97.86	**0.31**
ResNet50	93.86	92.34	87.74	97.45	97.90	0.59
DenseNet121	93.37	91.71	86.97	97.00	98.24	1.05
ResNet18 + CBAM	94.26	92.93	89.46	96.69	98.41	0.62
Proposed	**95.15**	**94.07**	**91.05**	97.30	**98.59**	0.68

Table 4 also reports the measured minibatch training time using four NVIDIA GeForce RTX 2080 Ti GPUs. According to the time column, Densenet 121 takes the most time to train among the listed models and is less memory efficient. Despite training two tasks, our approach is 35% quicker than the DenseNet121 model. For all other models, the better performing models have taken longer to train.

6.3. Ablation Studies

Since the attention supervision can be applied inside any residual stage (Figure 6), the first ablation study determines the best placement of the MGA module. The results in Table 5 indicate that the best performance is achieved by placing the MGA module at the third residual stage. One possible explanation from the perspective of multi-task learning is that increasing the number of shared hidden layers between the highly related tasks helps the performance Caruana (1997). Furthermore, branching out the attention supervision at the initial layers might not be sufficient for learning the image representation. Additionally, comparing the results in Table 5 with the other models' performance in Table 4 shows that using the MGA module in any location improves the overall classification performance.

Table 5. Comparison of different MGA module placements (the best performance in each metric is in red).

Method	Accuracy (%)	F1 score (%)	Recall (%)	Precision (%)	ROC AUC (%)	time (s)
First Residual Stage	94.65	93.49	90.86	96.29	98.08	**0.66**
Second Residual Stage	94.82	93.71	**91.25**	96.3	98.56	0.67
Third Residual Stage	**95.15**	**94.07**	91.05	**97.3**	**98.59**	0.68

The second experiment investigates the effectiveness of MGA classification for different training set sizes. We simplified the model and ran this experiment on a ResNet18 base model with only one embedded CBAM at the first residual stage to save the computation. Specifically, the single-task and multi-task models are identical, except that the spatial attention map of the CBAM is supervised with the predicted lesion masks for the multi-task learning case.

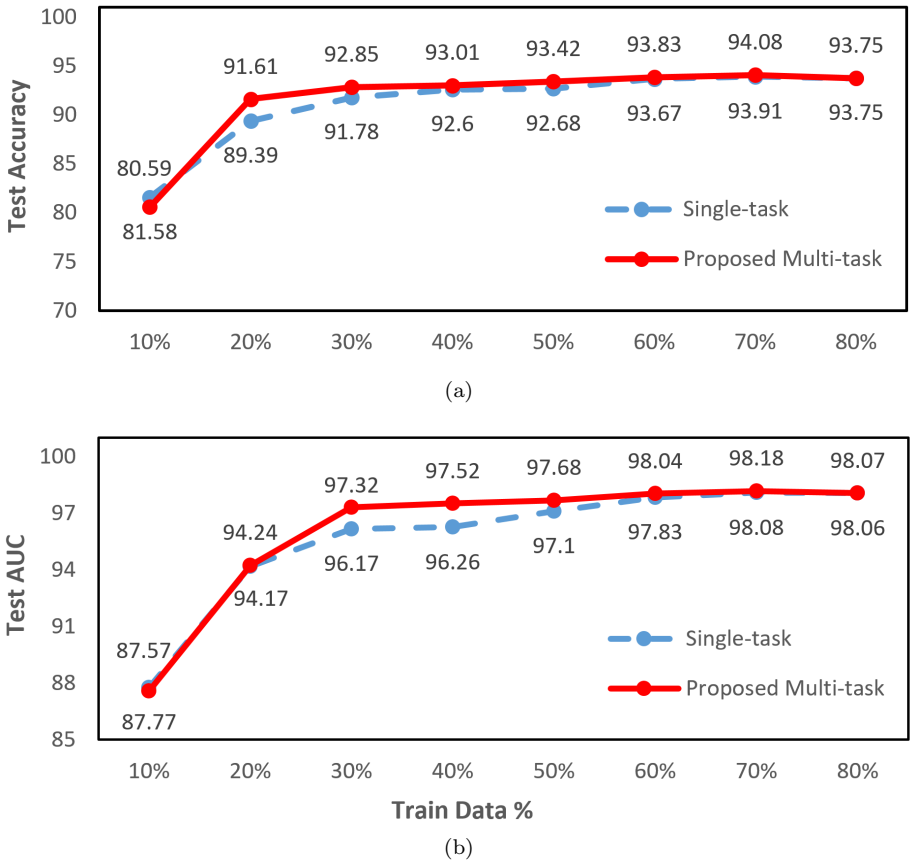


Figure 8. The Single-task vs. Multi-task classification performance of the simplified model measured by (a) Accuracy and (b) AUC on test set for different train set sizes

Figure 8 shows a test performance comparison of this single and multi-task classification for different train set sizes. While the test set is separated and fixed, the remaining data is split between the train and validation according to the train data size. It is worth noting that the percentages are not exact since the data should be divided in a patient-aware and stratified manner. Consistent with the literature Crichton et al. (2017); Gong et al. (2019), the results show that multi-task learning improves performance, especially when the training data is small and sufficient for the learning to happen. We can see that from 20% to 60% there is the most improvement. 10% is too small for learning, and for the large train sets, there is less difference in performance, yet the generalization and interpretability advantages remain. In other words, the proposed multi-task learning stands out in the model performance when the train set size

is sufficient but relatively small. Moreover, a 70-30 data split between the train and validation has given the best performance; therefore, it is the ratio we used for comparing all the models.

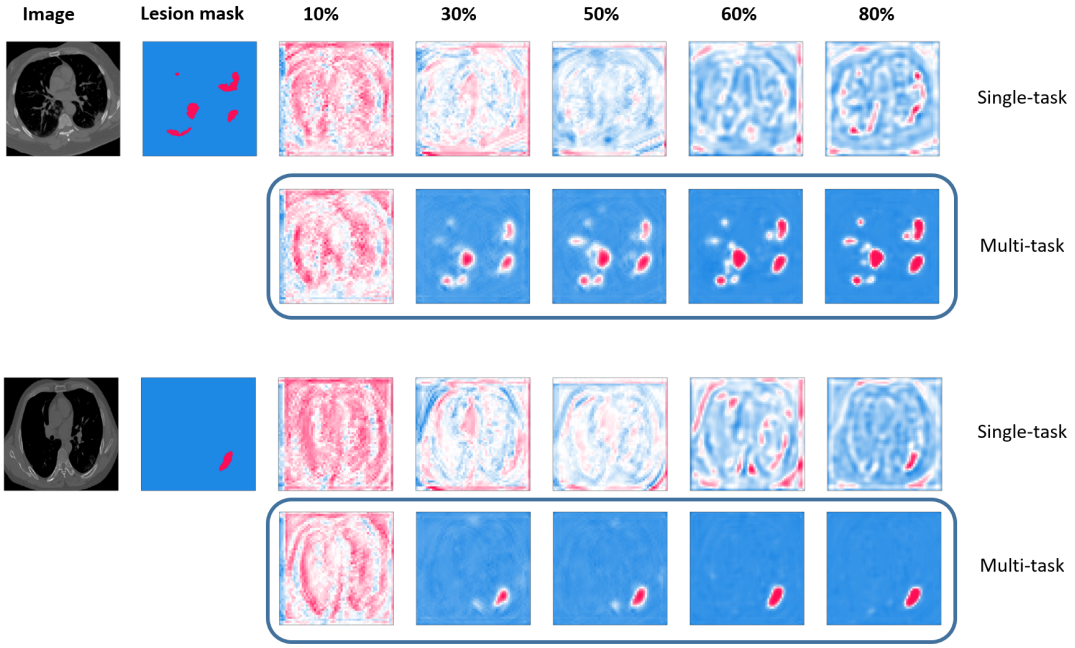


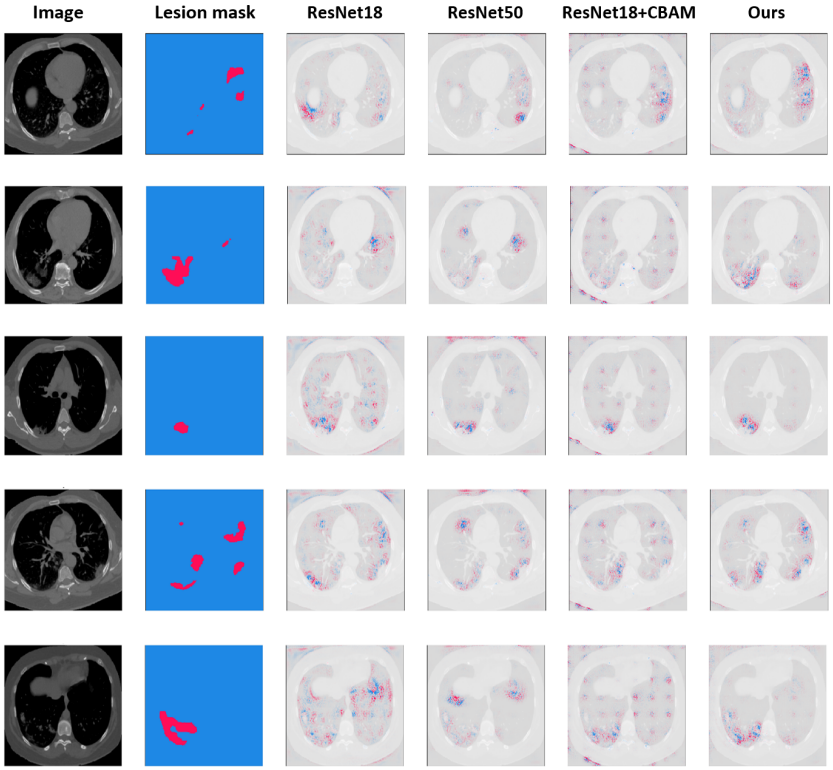
Figure 9. The Single-task vs. Multi-task attention maps of two example frames when different percentages of training data is used. The color changes from blue to red as the pixel's attention weight increases. While attention maps of both methods converge to highly score the lesions, the supervised attention maps in the multi-task classification converge using considerably less training data (20%). The unsupervised attention map takes a lot more data, 80% in our case, to focus on the lesions.

The attention maps depicted in Figure 9 reveal that, as the train data increases, although the unsupervised attention map of both the single-task learner and the multi-task attention map converge to the lesions, the latter one converges using only 30% of the data while the improved focus emerges in the former after using 80% of the data. Therefore, attention supervision can help the fast convergence of the attention map with a smaller required train data size.

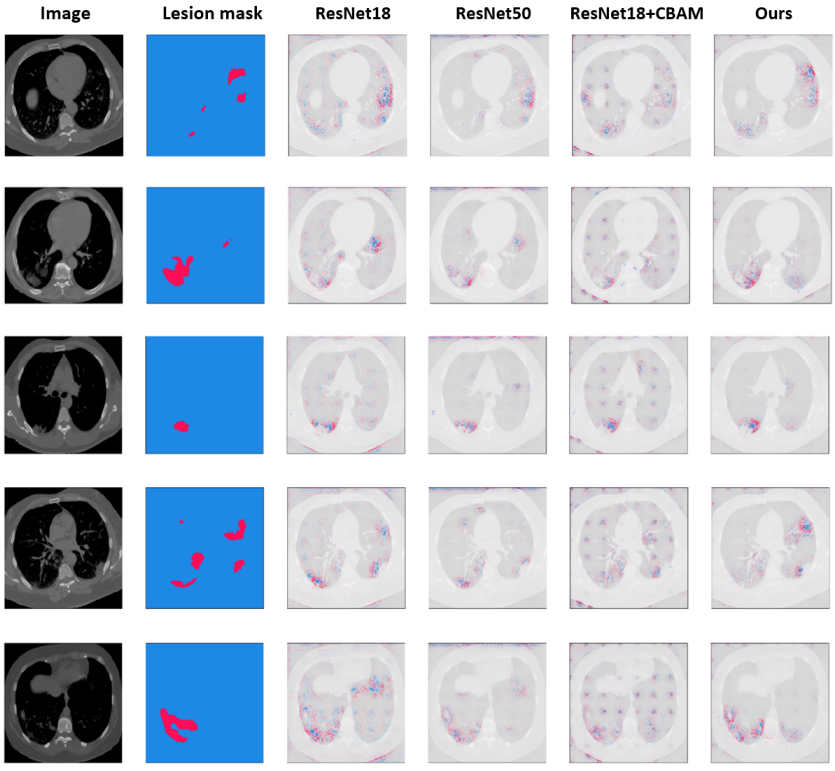
6.4. Interpretability using Attribution Maps

Figure 10 compares the attribution maps of our model with the other models for five COVID-19 frames, using 10a DeepLIFT (Rescale) Shrikumar et al. (2017) and 10b Integrated Gradient attribution Sundararajan et al. (2017) methods. We can see that the red and blue regions (pointing to influential features) of our model's attribution map highly overlap with the lesion regions (represented with red color in the lesion mask). In other words, the lesion regions highly contribute to our model decision while other models are less focused on the lesions. This visualization further emphasizes the effectiveness of our multi-task learning approach on improving the model's attention to the relevant regions. This is because, compared to the single-task (classification), the two integrated tasks (namely, attention supervision and classification)

can provide evidence for the relevance or irrelevance of specific features.



(a)



(b)

Figure 10. (a) DeepLIFT (Rescale) and (b) Integrated Gradient attribution comparison between different models.

Moreover, DeepLIFT (Rescale) and Integrated Gradients have generated highly correlated

attribution maps, consistent with the past works, while DeepLIFT is considerably faster to execute. Current attribution methods do not explain how the network combines the features to produce the answer and scores them independently, but DeepLIFT (RevealCancel) method takes dependencies into account. For future exploration, it would be interesting to derive and compare the DeepLIFT (RevealCancel) attribution maps, which claimed to outperform the two other techniques when Pytorch support is available.

7. Conclusion and Future Direction

This paper presented the MGA-based classification model, a novel multi-task learner for COVID-19 diagnosis based on CT scan images. Specifically, the proposed model leveraged the predicted lesion masks to impose extra supervision on the classifier's attention module. Since attention supervision and classification are consolidatory tasks, their multi-task learning yielded a significant performance improvement over the single-task baseline (i.e., the baseline model with attention) and the state-of-the-art deep learning methods in image classification. Our experiments also showed that the proposed method benefits from improved data efficiency and interpretability, which are especially valuable in the medical domain in which data may be often limited, and reliability is paramount. Additionally, in this work, a large, nationally diverse, and broadly representative COVID-19 CT slice classification dataset has been curated for conducting experiments and serving as a benchmark dataset for the research community. The quality of our dataset is ensured using slices with patient identification and precise labels.

This research could be extended to include an MGA module that segments both the lungs and the lesions to improve the overall inside lung learning, especially for normal cases. Additionally, as only two groups of COVID-19 and non-COVID-19 are examined in most of the literature, the effect of having more precisely categorized disease classes on COVID-19 detection could be further investigated.

Disclosure Statement

The authors report there are no competing interests to declare.

Data Availability Statement

The data Maftouni et al. (2021) that support the findings of this study are available in Kaggle at `https://www.kaggle.com/maedemaftouni/large-covid19-ct-slice-dataset`. These data were curated from the following resources available in the public domain: Afshar et al. (2021); Cohen et al. (2020); Jun et al. (2020); MedSeg (2020); Morozov et al. (2020); Rahimzadeh et al. (2021); Zhao et al. (2020).

References

Afshar, P., Heidarian, S., Enshaei, N., Naderkhani, F., Rafiee, M. J., Oikonomou, A., Fard, F. B., Samimi, K., Plataniotis, K. N., and Mohammadi, A. (2021). Covid-ct-md, covid-19 computed tomography scan dataset applicable in machine learning and deep learning. *Scientific Data*, 8(1):1–8.

Ai, T., Yang, Z., Hou, H., Zhan, C., Chen, C., Lv, W., Tao, Q., Sun, Z., and Xia, L. (2020). Correlation of chest ct and rt-pcr testing in coronavirus disease 2019 (covid-19) in china: a report of 1014 cases. *Radiology*, page 200642.

Amyar, A., Modzelewski, R., Li, H., and Ruan, S. (2020). Multi-task deep learning based ct imaging analysis for covid-19 pneumonia: Classification and segmentation. *Computers in Biology and Medicine*, 126:104037.

Badrinarayanan, V., Kendall, A., and Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bao, G., Chen, H., Liu, T., Gong, G., Yin, Y., Wang, L., and Wang, X. (2020). Covid-mtl: Multi-task learning with shift3d and random-weighted loss for automated diagnosis and severity assessment of covid-19. *arXiv preprint arXiv:2012.05509*.

Caruana, R. (1997). Multi-task learning. *Machine learning*, 28(1):41–75.

Chaganti, S., Grenier, P., Balachandran, A., Chabin, G., Cohen, S., Flohr, T., Georgescu, B., Grbic, S., Liu, S., Mellot, F., et al. (2020). Automated quantification of ct patterns associated with covid-19 from chest ct. *Radiology: Artificial Intelligence*, 2(4):e200048.

Chassagnon, G., Vakalopoulou, M., Battistella, E., Christodoulidis, S., Hoang-Thi, T.-N., Dangeard, S., Deutsch, E., Andre, F., Guillo, E., Halm, N., et al. (2020). Ai-driven ct-based quantification, staging and short-term outcome prediction of covid-19 pneumonia. *arXiv preprint arXiv:2004.12852*.

Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Cohen, J. P., Morrison, P., Dao, L., Roth, K., Duong, T. Q., and Ghassemi, M. (2020). Covid-19 image data collection: Prospective predictions are the future. *arXiv preprint arXiv:2006.11988*.

Crichton, G., Pyysalo, S., Chiu, B., and Korhonen, A. (2017). A neural network multi-task learning approach to biomedical named entity recognition. *BMC bioinformatics*, 18(1):1–14.

Fang, Y., Zhang, H., Xie, J., Lin, M., Ying, L., Pang, P., and Ji, W. (2020). Sensitivity of chest ct for covid-19: comparison to rt-pcr. *Radiology*, 296(2):E115–E117.

Gao, K., Su, J., Jiang, Z., Zeng, L.-L., Feng, Z., Shen, H., Rong, P., Xu, X., Qin, J., Yang, Y., et al.

(2021). Dual-branch combination network (dcn): Towards accurate diagnosis and lesion segmentation of covid-19 using ct images. *Medical image analysis*, 67:101836.

Goncharov, M., Pisov, M., Shevtsov, A., Shirokikh, B., Kurmukov, A., Blokhin, I., Chernina, V., Solovev, A., Gombolevskiy, V., Morozov, S., et al. (2021). Ct-based covid-19 triage: deep multi-task learning improves joint identification and severity quantification. *Medical image analysis*, 71:102054.

Gong, T., Lee, T., Stephenson, C., Renduchintala, V., Padhy, S., Ndirango, A., Keskin, G., and Elibol, O. H. (2019). A comparison of loss weighting strategies for multi task learning in deep neural networks. *IEEE Access*, 7:141627–141632.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

He, X., Yang, X., Zhang, S., Zhao, J., Zhang, Y., Xing, E., and Xie, P. (2020). Sample-efficient deep learning for covid-19 diagnosis based on ct scans. *medRxiv*.

Helwan, A., Ma'aitah, M. K. S., Hamdan, H., Ozsahin, D. U., and Tuncyurek, O. (2021). Radiologists versus deep convolutional neural networks: A comparative study for diagnosing covid-19. *Computational and Mathematical Methods in Medicine*, 2021.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. (2016). Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*.

Jun, M., Cheng, G., Yixin, W., Xingle, A., Jiantao, G., Ziqi, Y., Minqing, Z., Xin, L., Xueyuan, D., Shucheng, C., Hao, W., Sen, M., Xiaoyu, Y., Ziwei, N., Chen, L., Lu, T., Yuntao, Z., Qiongjie, Z., Guoqiang, D., and Jian, H. (2020). COVID-19 CT Lung and Infection Segmentation Dataset. https://doi.org/10.5281/zenodo.3757476. [Online].

Kendall, A., Gal, Y., and Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.

Lei, Q., Li, G., Ma, X., Tian, J., fan Wu, Y., Chen, H., Xu, W., Li, C., and Jiang, G. (2021). Correlation between ct findings and outcomes in 46 patients with coronavirus disease 2019. *Scientific Reports*, 11(1):1–6.

Lu, J., Yang, J., Batra, D., and Parikh, D. (2016). Hierarchical question-image co-attention for visual question answering. *arXiv preprint arXiv:1606.00061*.

Maftouni, M., Law, A. C. C., Shen, B., Grado, Z. J. K., Zhou, Y., and Yazdi, N. A. (2021). A robust ensemble-deep learning model for covid-19 diagnosis based on an integrated ct scan images database. In *IIE Annual Conference. Proceedings*, pages 632–637. Institute of Industrial and Systems Engineers

(IISE).

MedSeg (2020). COVID-19 CT segmentation dataset. http://medicalsegmentation.com/covid19/. [Online].

Misztal, K., Pocha, A., Durak-Kozica, M., Wator, M., Kubica-Misztal, A., and Hartel, M. (2020). The importance of standardisation–covid-19 ct & radiograph image data stock for deep learning purpose. *Computers in Biology and Medicine*, 127:104092.

Morozov, S., Andreychenko, A., Pavlov, N., Vladzymyrskyy, A., Ledikhova, N., Gombolevskiy, V., Blokhin, I. A., Gelezhe, P., Gonchar, A., and Chernina, V. Y. (2020). Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465*.

Nogueira, F. (2014). Bayesian Optimization: Open source constrained global optimization tool for Python.

Pang, Y., Xie, J., Khan, M. H., Anwer, R. M., Khan, F. S., and Shao, L. (2019). Mask-guided attention network for occluded pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4967–4975.

Pham, T. D. (2020). A comprehensive study on classification of covid-19 on computed tomography with pretrained convolutional neural networks. *Scientific reports*, 10(1):1–8.

Polsinelli, M., Cinque, L., and Placidi, G. (2020). A light cnn for detecting covid-19 from ct scans of the chest. *Pattern Recognition Letters*, 140:95–100.

Rahimzadeh, M., Attar, A., and Sakhaei, S. M. (2021). A fully automated deep learning-based network for detecting covid-19 from a new and large lung ct scan dataset. *Biomedical Signal Processing and Control*, page 102588.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Shamsi, A., Asgharnezhad, H., Jokandan, S. S., Khosravi, A., Kebria, P. M., Nahavandi, D., Nahavandi, S., and Srinivasan, D. (2021). An uncertainty-aware transfer learning-based framework for covid-19 diagnosis. *IEEE Transactions on Neural Networks and Learning Systems*.

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, C., Huang, Y., Ouyang, W., and Wang, L. (2018). Mask-guided contrastive attention model for

person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1179–1188.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR.

Tahan, S., Parikh, B. A., Droit, L., Wallace, M. A., Burnham, C.-A. D., and Wang, D. (2021). Sars-cov-2 e gene variant alters analytical sensitivity characteristics of viral detection using a commercial rt-pcr assay. *Journal of Clinical Microbiology*, pages JCM–00075.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR.

Tao, A., Sapra, K., and Catanzaro, B. (2020). Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*.

Tilborghs, S., Dirks, I., Fidon, L., Willems, S., Eelbode, T., Bertels, J., Ilsen, B., Brys, A., Dubbeldam, A., Buls, N., et al. (2020). Comparative study of deep learning methods for the automatic segmentation of lung, lesion and lesion type in ct scans of covid-19 patients. *arXiv preprint arXiv:2007.15546*.

Trivizakis, E., Tsiknakis, N., Vassalou, E. E., Papadakis, G. Z., Spandidos, D. A., Sarigiannis, D., Tsatsakis, A., Papanikolaou, N., Karantanas, A. H., and Marias, K. (2020). Advancing covid-19 differentiation with a robust preprocessing and integration of multi-institutional open-repository computer tomography datasets for deep learning analysis. *Experimental and therapeutic medicine*, 20(5):1–1.

Vakalopoulou, M., Chassagnon, G., Bus, N., Marini, R., Zacharaki, E. I., Revel, M.-P., and Paragios, N. (2018). Atlasnet: multi-atlas non-linear deep networks for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 658–666. Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., and Tang, X. (2017). Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Wang, J., Yu, X., and Gao, Y. (2021). Mask guided attention for fine-grained patchy image classification. *arXiv preprint arXiv:2102.02771*.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.

Wu, Y.-H., Gao, S.-H., Mei, J., Xu, J., Fan, D.-P., Zhang, R.-G., and Cheng, M.-M. (2021). Jcs: An explainable covid-19 diagnosis system by joint classification and segmentation. *IEEE Transactions on Image Processing*, 30:3113–3126.

Xie, X., Zhong, Z., Zhao, W., Zheng, C., Wang, F., and Liu, J. (2020). Chest ct for typical 2019-ncov

pneumonia: relationship to negative rt-pcr testing. *Radiology*, page 200343.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Yao, Q., Xiao, L., Liu, P., and Zhou, S. K. (2021). Label-free segmentation of covid-19 lesions in lung ct. *IEEE Transactions on Medical Imaging*.

Yazdani, S., Minaee, S., Kafieh, R., Saeedizadeh, N., and Sonka, M. (2020). Covid ct-net: Predicting covid-19 from chest ct images using attentional convolutional network. *arXiv preprint arXiv:2009.05096*.

Yuan, Y., Chen, X., and Wang, J. (2020). Object-contextual representations for semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 173–190. Springer.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

Zhang, K., Liu, X., Shen, J., Li, Z., Sang, Y., Wu, X., Zha, Y., Liang, W., Wang, C., Wang, K., et al. (2020). Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography. *Cell*.

Zhang, Y. and Yang, Q. (2017). A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*.

Zhao, J., Zhang, Y., He, X., and Xie, P. (2020). Covid-ct-dataset: a ct scan dataset about covid-19. *arXiv preprint arXiv:2003.13865*.