COVID-VIT: Classification of Covid-19 from 3D CT chest images based on vision transformer models

Xiaohong Gao [1], Maleika Khan[2], Rui Hui[3], Zhengmeng Tian,[3] Yu Qian[1], Alice Gao[4],

[1] Department of Computer Science, Middlesex University, London, UK.

x.gao@mdx.ac.uk

[3] University of Mauritius, Mauritius.

[3] General Hospital of PLA, Beijing, China.

[4] A&E Department, Newham University Hospital, Barts Health NHS Trust, London, UK.

alicegao@doctors.org.uk

Abstract

This paper aims to develop an explainable deep learning network to classify COVID from non-COVID based on 3D CT lung images. It applies a subset of the data for MIA-COV19 challenge through the development of 3D form of Vision Transformer deep learning architecture. The data comprise 1924 subjects with 851 being diagnosed with COVID, among them 1,552 being selected for training and 372 for testing. While most of the data volume are in axial view, there are a number of subjects' data are in coronal or sagittal views with 1 or 2 slices are in axial view. Hence, while 3D data based classification is investigated, in this competition, 2D axial-view images remains the main focus. Two deep learning methods are studied, which are vision transformer (ViT) based on attention models and DenseNet that is built upon conventional convolutional neural network (CNN). Initial evaluation results indicates that ViT performs better than DenseNet with F1 scores being 0.81 and 0.72 respectively. (Codes are available at GitHub at https://github.com/xiaohong1/COVID-ViT).

1. Introduction

COVID-19, officially known as SARS-CoV-2 is a strain of coronavirus. The first cases were seen in Wuhan, China~~,~~ in late December 2019 before spreading globally [1-3] which as and was classified as a pandemic in March 2020 [4]. At present there are more than 182 million people infected with the virus and 3.9 million of deaths [5] with new variants keep appearing.

The clinical picture can range from a mild common cold-like illness~~,~~ to a severe viral pneumonia leading to acute respiratory distress syndrome (ARDS) that is potentially fatal. The presence of COVID-19 in respiratory specimens was detected by next generation sequencing or real-time reverse transcription polymerase chain reaction (RT-PCR) methods, a laboratory technique combining reverse transcription of Ribonucleic acid (RNA) into Deoxyribonucleic acid (DNA) and amplification of specific DNA targets. While PCR tests offer many advantages, results are not usually available for at least several hours. On the other hand, high resolution Computerised Tomography (CT) are non-invasive, easy to operate and prevalent and hence can assist diagnosis for COVID-19 rapidly.

With regard to imaging features, it appears that bilateral infiltrates with peripheral opacities and patchy consolidation are the most common findings on chest radiographs (CXR) [6,7] and bilateral ground glass opacities is often a key finding on CT [8,9].

As confirmed cases continue~~s~~ to increase considerably all over the world, timely detection of the disease not only can provide supportive care required by patients but also can prevent further spread of the virus. Consequently, effective screening of infected patients appears to be a critical step in this fight against

COVID-19 as well as to circumvent the temporary shortage of RT-PCR kits to confirm COVID-19 infection.

The challenge here facing detecting COVID-19 based on chest CT images is that when the disease is at its early onset, the characteristic patterns present less obvious to the human eyes [10]. Hence, machine learning based approaches are applied to investigate COVID-specific biomarkers. In this study vision transformer architectures are investigated.

2. Related work

Vision transformer (ViT) have recently demonstrated its potentials in image processing by achieving comparable results while requiring fewer computational resources. Based on self-attention architectures, transformer becomes the leading model in natural language processing (NLP) [11]. For NLP, by employing attention models, i.e. transformers, training speed can be significantly improved hence enhancing the performance of neural machine translation applications. For image processing, vision transformers are emerging and starting to show protentials by applying to computer vision tasks, such as image recognition [12]. Specifically, ViT appears to demonstrate excellent performance when trained on sufficient data, outperforming a comparable state-of-the-art CNN with four times fewer computational resources.

One of the advantages that that Transformers present is computational efficiency and scalability. It has become possible to train models of unprecedented size, with over 100 billion parameters [13].

Figure 1 illustrates the architecture of ViT employed in this study. In this study, the ViT is implemented in pytorch and heavily based on the code at [14].

The training process takes place at a GPU sever that equipped with one Quadro RTX 8000 GPU and 64GB memory under Debian Linux operating system. While in the training, the 2D images are resized to 224×224×3 and 224×224×32 for 3D. For 3D training, each subject's 2D slices (in JPG format) are firstly converted into Analyze (7.5) format, with both header (.hdr) and image (.img) files. The patch size for the application of ViT model is 7×7 for 2D images and 8×8×8 for 3D volumes. It takes about 24 hours for training 80 epochs for 2D images and ~30 hours for 3D volumes.

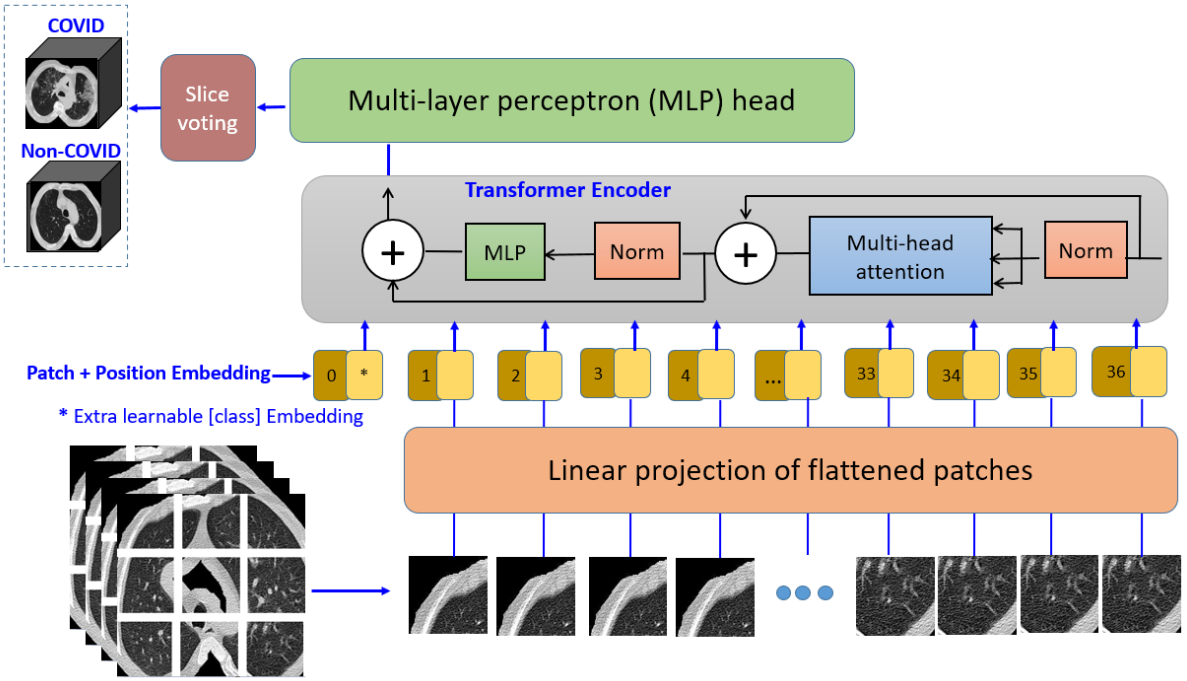

Figure 1. The 3D ViT architecture implemented in this work.

As illustrated in Figure 1, the classification of COVID applying a ViT architecture takes 8 steps, which are splitting an image into patches, flattening the patches, producing low-dimensional linear embeddings from the flattened patches, adding positional embeddings, inputting the sequence to a standard transformer encoder, pertaining the modelling with labels, and finetuning on the downstream datasets for image classification and finally, voting for image volume. In Figure 1, a volumetric image (x) in the space of $\mathbb{R}^{H\times W\times Z\times C}$ is reshaped into a sequence of flattened 2D patches $x_p \in \mathbb{R}^{N\times(P^3\cdot C)}$, where (H, W, Z) refers to the resolution (i.e. height, width, depth) of the original image volume whereas C the number of channels. At this study, $C = 1$ is for grey level images whilst (P, P, P) is for dimensions of patch-volume, leading $N = HWD/P^3$, the resulting number of patches.

Instead of using raw image patches, the input sequence is formed from feature maps extracted from a convolutional neural network (CNN) model. In this way, the patch embedding projection **E** (Eq. (1)) is employed to patches extracted from a CNN feature map. In this study, the feature map is extracted applying the built-in ViT extractor. The Transformer encoder comprises alternating layers of multi-headed self-attention (MSA) and Multilayer perception (MLP) blocks (Eqs. (2) & (3)). In addition, Layer normlisation (LN) is applied to every block and residual connections after every block.

In transformer encoder, MLP contains two layers with a Gelu non-linearity. Similar to Dosovitskiy et al. [12], the class token is prepended to sequence of embedded patches ($z_0^0 = x_{class}$), whose state that is the output of the Transformer encoder (z_L^0) serves as the image representation $\boldsymbol{y}$ (Eq.(4)).

$$\boldsymbol{z}_0 = \left[\boldsymbol{x}_{class;}\, \boldsymbol{x}_p^1\boldsymbol{E};\, \boldsymbol{x}_p^1\boldsymbol{E}; \ldots; x_p^1\boldsymbol{E}\right] + \boldsymbol{E}_{pos} \tag{1}$$

Where $\boldsymbol{E} \in \mathbb{R}^{D\times(P^3\cdot C)}$, $\boldsymbol{E_{pos}} \in \mathbb{R}^{(N+1)\times D}$

$$\boldsymbol{z}'_\ell = MSA(LN(\boldsymbol{z}_{\ell-1})) + \boldsymbol{z}_{\ell-1}, \quad \ell = 1 \ldots L \tag{2}$$

$$\boldsymbol{z}_\ell = MLP(LN(\boldsymbol{z}'_\ell)) + \boldsymbol{z}'_\ell, \qquad \ell = 1 \ldots L \tag{3}$$

$$\boldsymbol{y} = LN(\boldsymbol{z}_L^0) \tag{4}$$

3. The MIA-COV19 competition datasets

The CT thorax lung images are collected from MIA-COV19 competition [15-18]. Table 1 lists the detailed information the data applied in this paper. In total, the data from1926 subject are employed, consisting of 1,552 for training (20% for validation) and 372 for testing. The resolution of these images is either 512×512 or 768×768 pixels whereas the depth of each volume ranges from 4 slices to 1026.

Table 1. The datasets from MIA-COV19 competition applied in this paper. Note 2D slice numbers are the slices that have undertaken pre-processing stage and removed those with little lung contents.

Label	**Train**		**Testing**		**Total (subject)**
	3D subject (block)	(2D slice)	3D subject (block)	(2D slice)	
COVID	687 (11,490)	61,141	164 (2,388)	15,410	851
Non-COVID	865 (16,005)	82,197	208 (4,076)	19,681	1073
total	1,552 (27,495)	143,338	372 (6,464)	35,091	**1,924**

3.1 Image pre-processing

Because the diseased regions of a COVID-19 dataset occupy less than 10% of the whole volume and are presented in a certain number of slices, image pre-processing hence takes place first to maximise the large visibility of diseased slices while removing scanner artefact. Figure 2 demonstrates a montage view of a data set. It shows that the first 3 slices hardly depict any lung content whereas the boundary information as well as the background in each slice accommodates more than half of the slice in concern in each 2D image. In addition, the heart (arrow) and liver (arrow head) also make a large appearance in several slices.

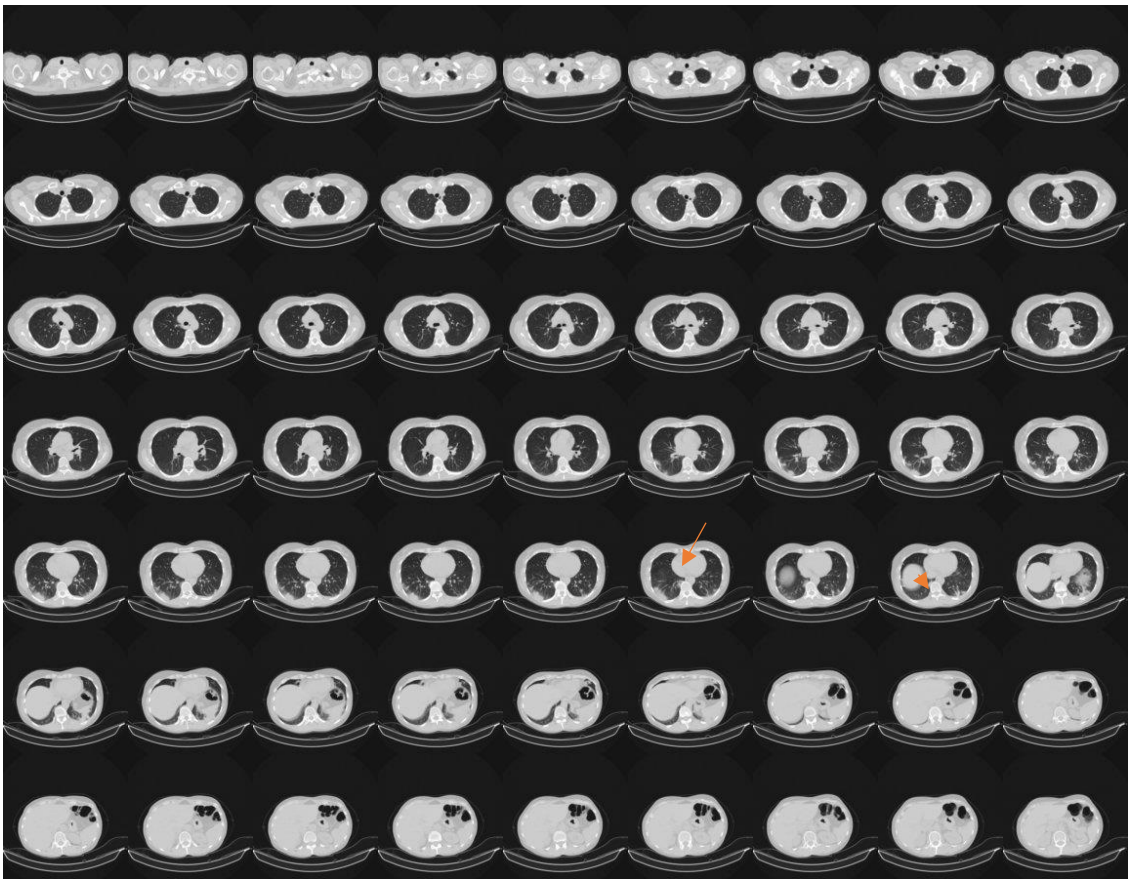

Figure 2. An axial view of a data volume in the form of montage. Arrow: heart. Arrow head: liver. This subject has confirmed diagnosis of COVID-19.

Hence, before the training and testing, all images undertake pre-processing stage to remove the boundary, which is illustrated in Figure 3(a). In addition, each dataset is divided into a number of each 3D blocks with a resolution of 224×224×16, which covers bottom (3(c)), left (3(b)) and right (3(d)) lung regions (pointed out by orange boxes in 3(a)) in the form of montage.

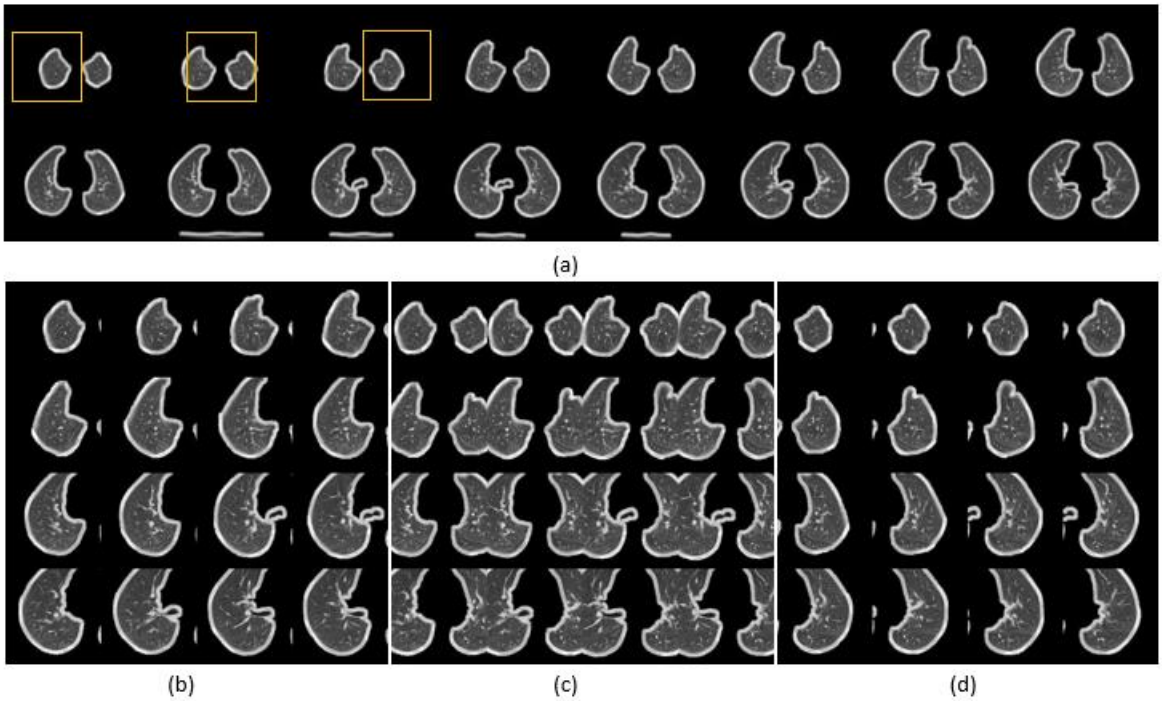

Figure 3. Generation of 3D blocks for each data set. (a) original image sequence after segmentation. (b) right lung segment montage, (c) Middle bottom region; (d) left lung region.

3.2 The challenges detecting COVID from Non-COVID CT images

Since there are still many unknowns regarding to COVID-19 features, many biomarkers attributed to COVID-19 are not specific. The common visible patterns of COVID-19 include bilateral involvement and peripheral distribution, with superimposed interlobular septal thickening and visible intralobular lines. However, other patterns, such as with uni-lobular, perihilar patchy ground glass distribution do exist with COVID-19 patients [19].

4. Experimental results

The classification results are subject based, which is calculated from the predicted scores of all the 2D or 3D components images for that subject. Considering the artefact that might be introduced during the pre-processing stage and not every slice of a COVID patient containing the disease features, the subject is classified as having COVID if more than a threshold (e.g. 25%) number of slices or 3D components are predicted as COVID. Similarly, if the remaining number, e.g. 75% or more, slices are predicted as nonCOVID, this subject will be classified as nonCOVID patient. Table 1 presents the confusion matrixes for the two deep learning systems, one is COVID-CT system based on DenseNet [20] and one is Vision Transformer architecture shown in Figure 1. The evaluation results are based on test dataset whilst the training data sets are divided into both training (80%) and validation (20%). It shows ViT performs slightly better than DenseNet with 78.8% accuracy in comparison of 76% for DenseNet.

Table 2. Evaluation results for the two deep leaning system, CNN-DenseNet and VIT, in the form of confusion matrix.

	CNN – DenseNet			**Vit—Vision transformer**		
	COVID	Non COVID	Average	COVID	Non COVID	Average
COVID (predict)	119	29	80.4	138	26	

Non COVID (predict)	64	130	67.0	44	164	
Sensitivity (%)	80.4	67.0	**72.8**	84.1	82.5	**83.3**
Specificity (%)	75.1	83.6	**79.3**	78.9	86.3	**82.6**
Accuracy (%)	77.1	74.9	**76**	84.1	78.8	**81.1**
F1-score	0.71	0.73	**0.72**	0.80	0.82	**0.81**

5. Conclusion

The aim of this work is to build an explainable system for medical application. Vision transformer (ViT) architectures are built upon attention models and are scalable when compared with CNN based models. Specifically, in the medical domain, the number of data sets can never be as large as current benchmark databases, e.g. ImageNet, with over millions of images. Hence a system that can still achieve good performance while employing limited number of datasets will make significant impact in the medical applications.

In comparison with CNN based model DenseNet for COVID-CT, ViT model appears to perform better with 78.8% accuracy whereas DenseNet realised accuracy of 76%.

While chest CT images are in 3D volume, it is a natural approach to process these data in 3D form. However, due to the large variations of slice numbers (depth), ranging from 4 to 1000+, with varying resolutions, generating 3D volumes present a challenge. Hence a depth of 16 slices, i.e. a volume of 224×224×16, is created for those subjects with sufficient depth images, by selecting slices evenly cross the whole volume. For example, if a 3D dataset has 64 slices in depth, then two sub-volumes are created for this subject with sub-volume 1 containing slices 1,3, … 31. and sub-volume 2 having slices of 2,4, …, 32.

As addressed previously, the lesioned regions are proportionally small in relation to the whole volume, which might constitute the main reason that 3D based system perform far worse (68% accuracy) than 2D based models (78.8%). Future work will further investigate this challenging issue. Another challenge remains to be the data volume size when preforming pre-pressing. Overall the training, validation and testing sizes are around 100 GB. Therefore pre-processing to segment lung content takes about 12 hours for all the subjects' dataset.

References:

[1] Zhu N, Zhang D, Wang W, Li, X, et al, A Novel Coronavirus from Patients with Pneumonia in China, 2019 , New England Journal of Medicine, 382(8):727-733, 2020.

[2] Perlman, S., Another Decade, Another Coronavirus. (2020) New England Journal of Medicine, 382(8): 760:762, 2020.

[3] Hui DS, I Azhar E, Madani TA, Ntoumi F, Kock R, et al, The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health - The latest 2019 novel coronavirus outbreak in Wuhan, China, International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases. 91: 264-266, 2020.

[4] WHO, Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020, Who.int. 2020. https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. Retrieved in April 2020.

[5] Resource Centre, Johns Hopkins University and Medicine, https://coronavirus.jhu.edu/.

[6] Chen N, Zhou M, Dong X, et al., Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study, The Lancet, 395 (10332):507:513, 2020.

[7] Huang, C., Wang, Y., Li, X., et al, Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China, Lancet, 395 (10223):497-506, 2020

[8] Raptis C, Hammer M, Short R, et al. Chest CT and Coronavirus Disease (COVID-19): A Critical Review of the Literature to Date, American Journal of Roentgenology, 1-4. 10.2214/AJR.20.23202, 2020.
[9] Kanne J, Little B, Chung J, Elicker B, Ketai L, Essentials for Radiologists on COVID-19: An Update—Radiology Scientific Expert Panel, Radiology, *in press*, 2020.
[10] Ng M, Lee E, Yang J, et al, Imaging Profile of the COVID-19 Infection: Radiologic Findings and Literature Review, Radiology: Cardiothoracic Imaging, 2:1, 2020.
[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In NIPS, 2017.
[12] Dosovitskiy A., et al., An image is worth 16x16 words: transformers for image recognition at scale, ICLR 2021.
[13] Brown TB, Mann B, Ryder N, Subbiah M, et al., Language models are few-shot learners. arXiv, 2020.
[14] ViT implementation, https://github.com/lucidrains/vit-pytorch.
[15] Kollias D, Arsenos A, Soukissian L, Kollias S, MIA-COV19D: COVID-19 Detection through 3-D Chest CT Image Analysis, 2021**,** arXiv preprint arXiv:2106.07524, 2021.
[16] D. Kollias, et. al., Deep transparent prediction through latent representation analysis, 2020, arXiv preprint arXiv:2009.07044, 2020.
[17] D. Kollias, et. al., Transparent Adaptation in Deep Medical Image Diagnosis, TAILOR, pp251-267, 2020.
[18] D. Kollias, et. al.: "Deep neural architectures for prediction in healthcare", 2018, Complex & Intelligent Systems, 4(2):119-131, 2018.
[19] Tran TA, Cezar R, Frandon J, et al., CT scan does not make a diagnosis od Covid-19: a cautionary case report, International Journal of indfectious diseases, 100: 182-183, 2020.
[20] Zhao J, Zhang Y , He,X , Xie, P, COVID-CT-Dataset: a CT scan dataset about COVID-19, arXiv preprint arXiv:2003.13865, 2020.