Advancing COVID-19 Detection in 3D CT Scans

Qingqiu Li[1], Runtian Yuan[2], Junlin Hou[3], Jilan Xu[2], Yuejie Zhang[2*], Rui Feng[2*], Hao Chen[3*]

[1] School of Academy for Engineering and Technology, Fudan University, China
[2] School of Computer Science, Fudan University, China [3] Department of Computer Science
and Engineering, The Hong Kong University of Science and Technology, China
{qqli22,rtyuan21,jilanxu18,yjzhang,fengrui}@fudan.edu.cn,csejlhou@ust.hk,jhc@cse.ust.hk

Abstract

To make a more accurate diagnosis of COVID-19, we propose a straightforward yet effective model. Firstly, we analyze the characteristics of 3D CT scans and remove the non-lung parts, facilitating the model to focus on lesion-related areas and reducing computational cost. We use ResNeSt-50 as the strong feature extractor, exploring various pre-trained weights and fine-tuning methods. After a thorough comparison, we initialize our model with CMC v1 pre-trained weights which incorporate COVID-19-specific prior knowledge, and perform Visual Prompt Tuning to reduce the number of training parameters. The superiority of our model is demonstrated through extensive experiments, showing significant improvements in COVID-19 detection performance compared to the baseline model. Among 12 participating teams, our method ranked 4th in the 4th COVID-19 Competition Challenge I with an average Macro F1 Score of 94.24%.

1. Introduction

The outbreak of COVID-19 has led to widespread health crises and fatalities. Early detection is crucial for controlling and preventing the spread of the virus. As shown in Fig. 1, Chest CT scans have been extensively utilized for diagnosing and monitoring COVID-19 patients, due to their ability to provide detailed insights into lung involvement's extent and severity. However, the vast number of CT images generated necessitates a significant workload for radiologists and medical practitioners, making the diagnosis process challenging.

In recent years, deep learning has been widely applied in the automatic detection of COVID-19 [7, 9, 19]. Xu et al. [29] utilized a 3D deep learning model to identify potential infection areas in CT scans and built upon this with a location-attention model for accurate COVID-19 detection. Kolliaz et al. introduced the COV19-CT-DB dataset [16, 17], containing a large volume of labeled data for both COVID-19 and non-COVID-19 cases. This initiative significantly advances the field by addressing the pressing need for comprehensive datasets, which are vital for the application of deep learning in the fight against COVID-19. Hou et al. [6] developed a methodology for diagnosing COVID-19 that leverages contrastive representation learning to capture the intra-class similarity and inter-class difference. They also implemented an adaptive joint training strategy combining multiple types of losses, including classification, mixup, and contrastive losses, to refine the learning process. However, COVID-19 is a novel disease that requires models to have a high degree of specialized medical knowledge. Moreover, computer-aided diagnostic systems place high demands on the real-time capabilities and computational efficiency of models. These factors present challenges for current COVID-19 diagnostics.

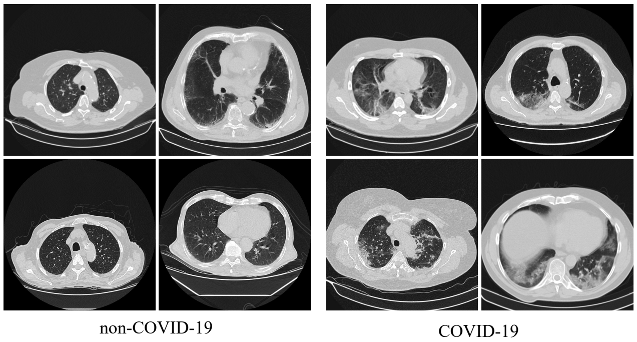

Figure 1. Samples of non-COVID-19 and COVID-19 from the COV19-CT-DB database.

To address the issues mentioned, we first preprocess 3D CT scans by removing non-lung slices that are irrelevant to the diagnosis of COVID-19. This encourages the model to concentrate on lesion-related areas and reduces computa-

[*] Corresponding author

tional cost to some extent. We employ ResNeSt-50 as the strong feature extractor. Considering that training a model from scratch leads to poor results, we attempt three different pre-trained weights, i.e., ImageNet [5], MIS-FM [27] and CMC v1 [6]. ImageNet is based on natural images; MIS-FM is trained on CT scans of various human body parts, i.e., neck and abdomen; CMC v1 is derived from chest CT scans, enriched with COVID-19-specific prior knowledge, providing a valuable foundation for models targeting this novel disease. Furthermore, given the real-time performance and computational demands of computer-aided diagnostic systems, we experiment with three distinct fine-tuning approaches, i.e., Full Fine-tuning, Linear Classification and Visual Prompt Tuning. Full Fine-tuning offers the highest accuracy while at the cost of increased parameters. Linear Classification significantly reduces the number of parameters during training, but it also leads to a considerable decrease in performance. Visual Prompt Tuning manages to reduce the number of required parameters with only a slight compromise on accuracy, meeting our criteria for effectiveness and efficiency.

Our primary contributions are outlined as follows:

1. We analyze the characteristics of 3D CT scans and remove the non-lung parts, facilitating the model to focus on lesion-related areas and reducing computational cost.
2. We utilize ResNeSt-50 as a strong feature extractor, comparing various pre-trained weights and fine-tuning methods, aiming to achieve a balance between effectiveness and efficiency.
3. Based on the CMC v1 pre-trained weight and the Visual Prompt Tuning, we achieve a Macro F1 score of 93.55% on the validation set of Challenge I, surpassing the baseline by 15.55%, while reducing the number of training parameters to 1.03M, which is 1/50 of the Full Fine-tuning method.

2. Related Work

2.1. COVID-19 Detection

In recent years, numerous advanced methods have emerged for COVID-19 detection. Song et al. [24] developed a deep learning-based CT diagnostic system that can accurately identify COVID-19 patients, distinguishing them from 100 bacterial pneumonia cases and 86 healthy individuals. Xu et al. [29] utilized a 3D deep learning model to segment potential infection areas from lung CT scans and introduced a location-attention model to differentiate between COVID-19, Influenza-A viral pneumonia, and healthy cases. Arsenos et al. [1] introduced the COV19-CT-DB dataset, containing a large volume of labeled data for both COVID-19 and non-COVID-19 cases, offering a solid foundation for enhancing model performance. Furthermore, they developed a CNN-RNN based classification model for COVID-19 detection, leveraging the strengths of both convolutional and recurrent neural networks to effectively handle the complexities of COVID-19 detection in CT scans. Hsu et al. [10] proposed a slice selection method for each CT dataset to filter out uncertain slices and a spatial-slice feature learning technique that employs a conventional and efficient backbone model for slice feature training. Hou et al. [6] designed a COVID-19 diagnosis approach with contrastive representation learning to effectively capture the intra-class similarity and inter-class difference. Besides, an adaptive joint training strategy was used to integrate classification loss, mixup loss, and contrastive loss.

2.2. Transfer Learning

As neural networks grow deeper and the number of parameters increases, transfer learning has become a pivotal study area for vision tasks. Transfer learning of pre-trained models can be categorized into three main types: Full Fine-tuning, Head-oriented, and Backbone-oriented. Full Fine-tuning means updating all backbone and classification head parameters of the pre-trained model. Although this method can achieve high accuracy, it involves a large number of parameters, and each single task has its own unique set of parameter weights. To address this, a popular method is to only fine-tune a subset of parameters, typically the classifier head, known as Head-oriented tuning [4, 11, 21]. Additionally, Backbone-oriented approaches [22, 31] fine-tune the pre-trained model by adding extra residual blocks or adapters to the backbone, offering an alternative strategy for model adaptation. The methods mentioned above fine-tune the pre-trained model itself, and some approaches [12, 32] choose to adjust the input. Jia et al. [12] proposed Visual Prompt Tuning to modify the input to the model, which introduces a small amount of task-specific learnable parameters into the input space while freezing the entire pre-trained backbone during downstream training, achieving good results with minimal parameters. Visual Prompt Tuning has also been applied in the medical field, Zhang et al. [32] enhanced it by incorporating a wide range of biomedical textual prompts, leading to adapt foundation models towards pathological image understanding.

The potential of transfer learning on COVID-19 detection has also been explored in some research works. For example, Subramanian et al. [25] used models pre-trained on ImageNet [5] and introduced the Learning without Forgetting (LwF) framework to boost the model's generalization across both known and new data. Li et al. [20] introduced a model pre-trained on the Chest X-ray14 dataset [28], enabling it to distinguish COVID-19 samples by leveraging its existing understanding of conventional pneumonia. Hou et al. [8] incorporated 3D weights pre-trained on video datasets, i.e. k400 [3], equipping the model with the capability to capture temporal information in 3D data.

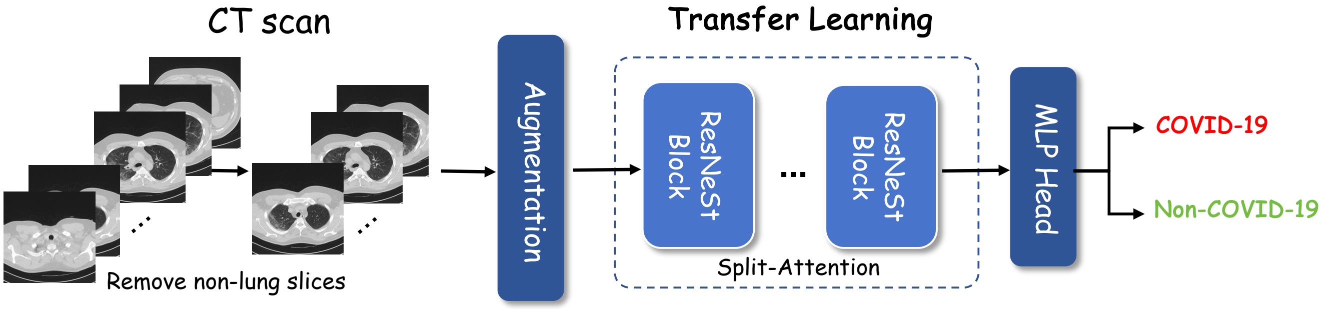

Figure 2. Overview of our framework for COVID-19 detection.

The aforementioned methods have achieved promising results in COVID-19 detection, but they still face some issues. The introduced pre-trained weights lack medical domain knowledge and are unaware of the novel disease COVID-19. Additionally, they all involve Full Fine-tuning, resulting in a large number of parameters and lengthy training times. Based on this, we explore more suitable pre-training weights and fine-tuning methods that simultaneously consider both effectiveness and efficiency, aiming to advance COVID-19 detection in 3D CT scans.

3. Methodology

The overall framework of our model is shown in Fig. 2. Given a minibatch of N randomly sampled CT scans and their pneumonia-type labels $\{(x_i, y_i)\}_{i=1,...,N}$, we first preprocess the 3D CT scans by removing slices unrelated to the lungs. Specifically, we find that each 3D CT volume contains some slices that do not contribute to COVID-19 detection, i.e., the neck area at the start and the abdominal region towards the end. Therefore, we remove the first 15% and the last 15% of slices from each $x_i \in \mathbb{R}^{1\times D\times H\times W}$ to obtain $\tilde{x}_i \in \mathbb{R}^{1\times D'\times H\times W}$, directing the model's focus towards areas relevant to lesion detection and simultaneously reducing computational cost. Here, H, W denote the height, and width of a CT slice, respectively. D represents the original number of slices, and D' denotes the quantity after pre-processing.

After obtaining the $\tilde{x}_i$, we utilize ResNeSt-50 [30] as our feature extractor, which presents a modular split-attention block within the individual network blocks to enable attention across feature-map groups. Considering that training a model from scratch leads to poor results, we employ transfer learning. We attempt three different pre-trained weights and three different fine-tuning methods, aiming to advance more effective and efficient COVID-19 detection.

We utilize three distinct pre-trained weights, each serving a unique purpose. **(1) ImageNet** [5]. In transfer learning, it is a common practice to initialize the model on downstream tasks with weights pre-trained on a large-scale ImageNet dataset. ImageNet is good for learning general image features, i.e., color and shape. **(2) MIS-FM** [27]. We adopt MIS-FM, which is trained on three different scales of CT datasets, including CT scans from various parts of the human body, i.e., neck, abdomen and lungs. The pre-trained weights provided by MIS-FM incorporate medical knowledge relevant to CT imaging and are beneficial for compensating the limited training data. **(3) CMC v1** [6]. We introduce the pre-trained weights based on CMC v1, specialized for lung characteristics and including COVID-19-specific prior knowledge.

We also employ different fine-tuning paradigms. The orange and blue colors represent learnable and frozen parameters, respectively.

Full Fine-tuning. As shown in Fig. 3(a), during Full Fine-tuning, the entire backbone and head are updated:

$$r_i = \text{Backbone}(\tilde{x}_i), \tag{1}$$

$$\tilde{y}_i = \text{Head}(r_i), \tag{2}$$

where $r_i \in \mathbb{R}^{d_e}$ is the representation vector in the d_e-dimensional latent space, and $\tilde{y}_i$ is the predicted probability of the sample $\tilde{x}_i$.

Linear Classification. As shown in Fig. 3(b), we adopt the Head-oriented approach, where only the classification head is trained:

$$r_i = \text{Backbone}(\tilde{x}_i), \tag{3}$$

$$\tilde{y}_i = \text{Head}(r_i). \tag{4}$$

Visual Prompt Tuning. As shown in Fig. 3(c), we explore the Visual Prompt Tuning approach, where a trainable vector P is concatenated onto the input $\tilde{x}_i$:

$$r_i = \text{Backbone}([\tilde{x}_i, P]), \tag{5}$$

$$\tilde{y}_i = \text{Head}(r_i). \tag{6}$$

We consider two concatenation methods: *below* and *pad*. *Below* means concatenating P and $\tilde{x}_i$ along the channel dimension, resulting in $\tilde{x}_i' \in \mathbb{R}^{(1+l)\times D'\times H\times W}$. *Pad* means

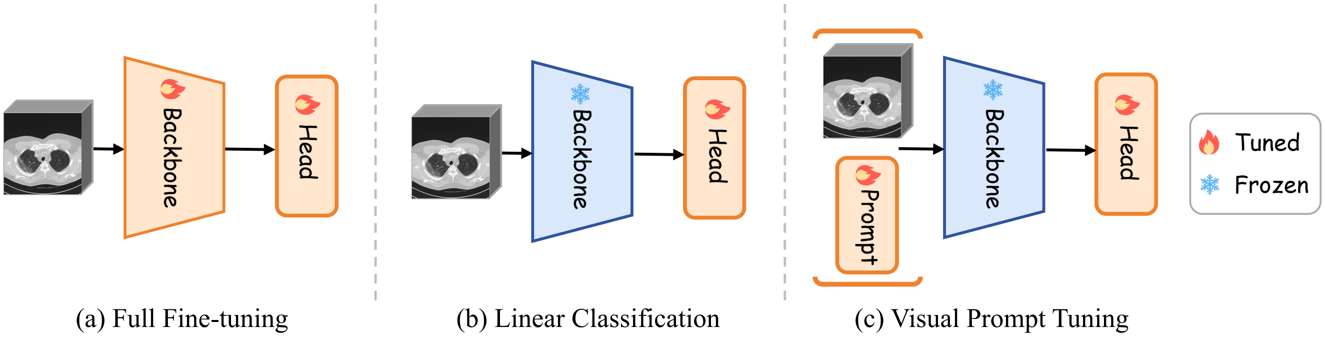

Figure 3. Different fine-tuning paradigms.

Table 1. Statistics of the Challenge I dataset.

Split	COVID-19	Non-COVID-19	Total
Training	703	655	1358
Validation	156	170	326
Testing	-	-	1413

concatenating P and $\tilde{x}_i$ along the D, H, and W dimensions, resulting in $\tilde{x}_i' \in \mathbb{R}^{1\times(D'+l)\times(H+l)\times(W+l)}$. Here, l denotes the token length of prompt vector P.

Finally, the standard cross-entropy loss is utilized for binary classification training, which is defined as:

$$L_{cls} = \frac{1}{N}\sum_{i=1}^{N} L_{ce}^{i}, \tag{7}$$

$$L_{ce}^{i} = -y_i^{\top} \log \tilde{y}_i, \tag{8}$$

where y_i denotes the one-hot vector of ground truth label.

4. Datasets

We evaluate our proposed approach on the COV19-CT-DB database [19]. The COV19-CT-DB contains chest CT scans, collected in various medical centers. The database includes 7,756 3D CT scans, where 1,661 are COVID-19 samples, whilst 6,095 refer to non-COVID-19 ones. In total, 724,273 slices correspond to the CT scans of the COVID-19 category and 1,775,727 slices correspond to the non-COVID-19 category class [1, 2, 14–18].

For Challenge I, the training set contains 1358 3D CT scans (655 non-COVID-19 cases and 703 COVID-19 cases). Based on this, we further enrich our training set with annotated data from Challenge II (120 non-COVID-19 cases and 120 COVID-19 cases), aiming to enhance the model's learning capacity and its ability to generalize across diverse cases of COVID-19 detection. The validation set consists of 326 3D CT scans (170 non-COVID-19 cases and 156 COVID-19 cases). The testing set includes 1413 scans and the labels are not available during the challenge.

5. Experiments

5.1. Data Pre-Processing

Our data pre-processing procedure is as follows. All 2D chest CT scan series are composed into a 3D volume of shape (D, H, W), where D, H, W denotes the number of slice, height, and width, respectively. Then, each 3D volume is resized to dimensions of (128, 256, 256). Finally, we transform the CT volume to the interval [0, 1] for intensity normalization.

5.2. Implementation Details

We utilize 3D ResNeSt-50 as the backbone of our model. For training, data augmentations include random resized cropping on the transverse plane, random cropping on the vertical section to 64, rotation, and color jittering. We use Adam algorithm [13] as our optimizer, setting the learning rate to $1e-4$ and the weight decay to $1e-5$. Our model is trained 100 epochs on 4 RTX 3090 GPUs with a batch size of 2 per GPU. When doing Visual Prompt Tuning, the token length of prompt vector P is set to 5.

5.3. Evaluation Metrics

Macro F1 score is adopted as the evaluation metric, which calculates the F1 score for each category separately and then averages these scores to assess overall performance. The F1 score for the i-th class is defined as:

$$\text{F1-Score}_i = 2 \times \frac{\text{Recall}_i \times \text{Precision}_i}{\text{Recall}_i + \text{Precision}_i}. \tag{9}$$

The Macro F1 score is an average of the F1 Score for COVID-19 and the F1 Score for non-COVID-19, which can be formulated as:

$$\text{Macro-F1} = \frac{1}{2}(\text{F1-Score}_{c19} + \text{F1-Score}_{n-c19}). \tag{10}$$

5.4. Ablation Study

Impact of different pre-trained weights. As shown in Table 2, we conduct transfer learning on three different pre-trained weights, corresponding to IDs 2-4. Among them,

Table 2. The results on the validation set of COVID-19 detection challenge.

ID	Method	Pre-trained	Params	Accuracy	Macro F1	F1	
						Non-COVID-19	COVID-19
	Full Fine-tuning						
1	Baseline	-	-	-	78.00	-	-
2	ResNeSt-50	ImageNet	50.40M	88.34	88.34	88.42	88.27
3	PCT-Net	MIS-FM	43.22M	88.96	88.93	88.39	89.47
4	ResNeSt-50	CMC v1	50.40M	93.87	93.86	93.63	94.08
5	ResNeSt-50 + mix	CMC v1	50.40M	93.25	93.25	93.08	93.41
6	ResNeSt-50 + mix + con	CMC v1	54.65M	93.87	93.86	93.71	94.01
	Linear Classification						
7	ResNeSt-50	CMC v1	0.003M	92.64	92.64	92.55	92.73
	Visual Prompt Tuning						
8	ResNeSt-50 + below	CMC v1	20.04M	93.56	93.56	93.42	93.69
9	ResNeSt-50 + pad	CMC v1	1.030M	93.56	93.55	93.29	93.80

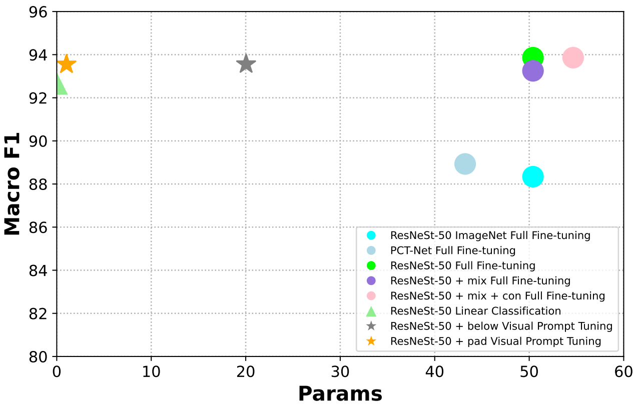

Figure 4. Overall performance comparison.

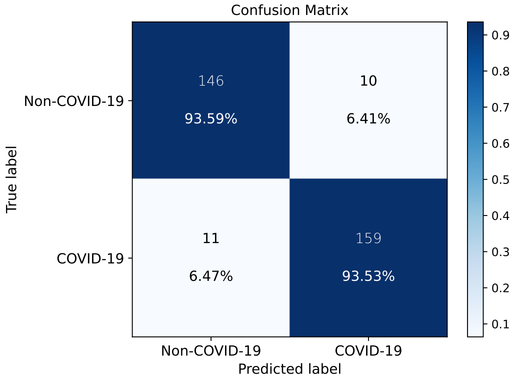

Figure 5. The confusion matrix of ID 9 model's prediction.

fine-tuning based on ImageNet weights yields the lowest results due to its lack of knowledge in the medical domain. Fine-tuning with MIS-FM weights results in a relatively minor increase of 0.5% in Macro F1 score compared to the former. Note that, as MIS-FM is trained based on PCT-Net [27], we use PCT-Net as the backbone. Fine-tuning with CMC v1 weights provides a significant boost to the model, achieving 93.86% on Macro F1 score, which is more than 5% higher than the previous two. This indicates that the COVID-19-specific knowledge introduced in CMC v1 is highly beneficial for the model to learn lesion features and accurately classify them.

Impact of different fine-tune methods. Focusing on the results of IDs 4, 7-9 in Table 2, ID 4 fine-tunes the entire backbone and head, achieving the highest Macro F1 score of 93.86%. However, it also has the highest number of training parameters, totaling 50.40M. On the other hand, ID 7 utilizes the Linear Classification strategy, training only the head, resulting in a significantly smaller number of parameters, almost close to zero. However, this approach also leads to some performance decrease, with a 1.2% decrease in Macro F1 score compared to ID 4. Meanwhile, IDs 8-9 employ the Visual Prompt Tuning approach to balance model performance and efficiency. In this approach, *below* indicates that the trainable prompt vector P is concatenated along the channel dimension of the input $\tilde{x}_i$, while *pad* means the prompt is wrapped around the input $\tilde{x}_i$. Based on *pad*-based Visual Prompt Tuning, only 1/50 of the parameters of Full Fine-tuning are utilized to achieve comparable results, with only a 0.3% decrease in Macro F1. Therefore, we consider the Visual Prompt Tuning-based transfer learning approach suitable for COVID-19 detection tasks that require both accuracy and speed. Fig. 5 shows the confusion matrices for the *pad*-based Visual Prompt Tuning method.

Impact of other techniques. In addition to experimenting with different pre-trained weights and fine-tuning methods, we explore some techniques that have been proven effective in previous work [6, 8]. As shown in IDs 5-6 in Table 2, *mix* represents the data augmentation method mixup, and *con* represents the contrastive learning method for COVID-19

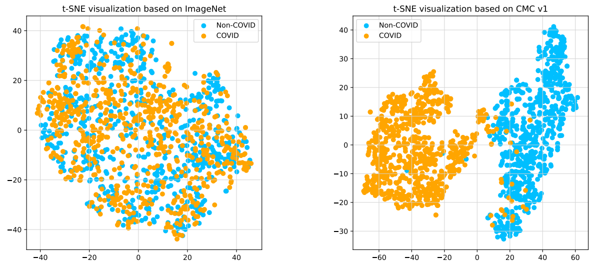

Figure 6. The t-SNE visualizations of encoded image representations on the training set.

detection introduced by Hou et al. [6]. In our experiments, the incorporation of these two techniques does not lead to an increase in Macro F1 score. This is likely that the primary motivation behind these techniques is to assist the model in more effectively learning the characteristics of COVID-19 lesions when data availability is limited. However, the CMC v1 pre-trained weights we introduced are robust and have alleviated this issue to a certain extent.

Fig. 4 illustrates the overall performance comparison of these methods, highlighting the superior performance achieved by Visual Prompt Tuning based on the pre-trained CMC v1 weights in both effectiveness and efficiency.

5.5. Visualization Results

We present the t-SNE [26] plots in Fig. 6 to visualize the feature embeddings of the images of Challenge I. The left plots feature embeddings of the images encoded with ImageNet weights, while the right utilizes CMC v1 weights. Notably, CMC v1 effectively clusters different categories, introducing robust prior knowledge to assist our model in learning meaningful disease-level semantic information for COVID-19 and non-COVID-19.

To further demonstrate the model's mechanism, we employed the Gradient-weighted Class Activation Mapping (Grad-CAM [23]) method to create heatmaps which highlight areas the model focuses on. As shown in Fig. 7, selecting three COVID-19 CT scans from the COV19-CT-DB dataset's validation set, the model successfully highlights the COVID-19 lesion areas in the lungs, suggesting clear interpretability of our model's diagnostic results. These heatmaps could potentially serve as a basis for COVID-19 diagnosis in clinical practice.

5.6. Results on Challenge I Leaderboard

Table 3 shows the results of our method and other participants on the testing set of the 4th COVID-19 detection challenge. Our team secures the 4th position, with a 0.65% difference in Macro F1 score from the 1st place. In our future work, we aim to focus on more refined data preprocessing and adopting a more effective method for lung region extraction. Moreover, we will explore stronger backbones and training strategies to enhance our model's performance.

Table 3. The competition results on the testing set of the COV19-CT-DB database.

Rank	Teams	Macro F1	F1(NC)	F1(C)
1	MDAP	94.89	95.97	93.81
2	Deep-Adaptation	94.60	95.53	93.66
3	ACVLAB	94.39	95.52	93.26
4	FDVTS (Ours)	94.24	95.41	93.07
5	ViGIR Lab	93.63	94.97	92.29
6	M2@Purdue	90.14	92.06	88.22
7	baseline	85.11	87.48	82.74

6. Conclusion

In this paper, we propose a straightforward yet effective model for COVID-19 detection. Firstly, we analyze the characteristics of 3D CT scans, removing non-lung regions from the entire volume. This approach not only facilitates the model to focus on lesion-related areas but also reduces computational cost. We choose ResNeSt-50 as the feature extractor, utilizing transfer learning instead of training the model from scratch. We initialize our model with pre-trained weights from CMC v1 to incorporate COVID-19-specific prior knowledge. Based on the aforementioned techniques, our model achieves a Macro F1 score of 93.55% on the validation set of Challenge I, surpassing the baseline by 15.55%, while reducing the number of training parameters to 1.03M, which is 1/50 of the full fine-tune method.

Acknowledgement

This work was supported by National Science and Technology Major Project (No. 2021ZD0114001), Natural Science Foundation of China (No. 62172101), the Science and Technology Commission of Shanghai Municipality (No. 21511101000; No. 23511100602), and the Hong Kong Innovation and Technology Fund (Project No. ITS/028/21FP). Yuejie Zhang, Rui Feng and Hao Chen are corresponding authors.

References

[1] Anastasios Arsenos, Dimitrios Kollias, and Stefanos Kollias. A large imaging database and novel deep neural architecture for covid-19 diagnosis. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*, pages 1–5. IEEE, 2022. 2, 4

[2] Anastasios Arsenos, Andjoli Davidhi, Dimitrios Kollias, Panos Prassopoulos, and Stefanos Kollias. Data-driven covid-19 detection through medical imaging. In *2023 IEEE International Conference on Acoustics, Speech, and Signal*

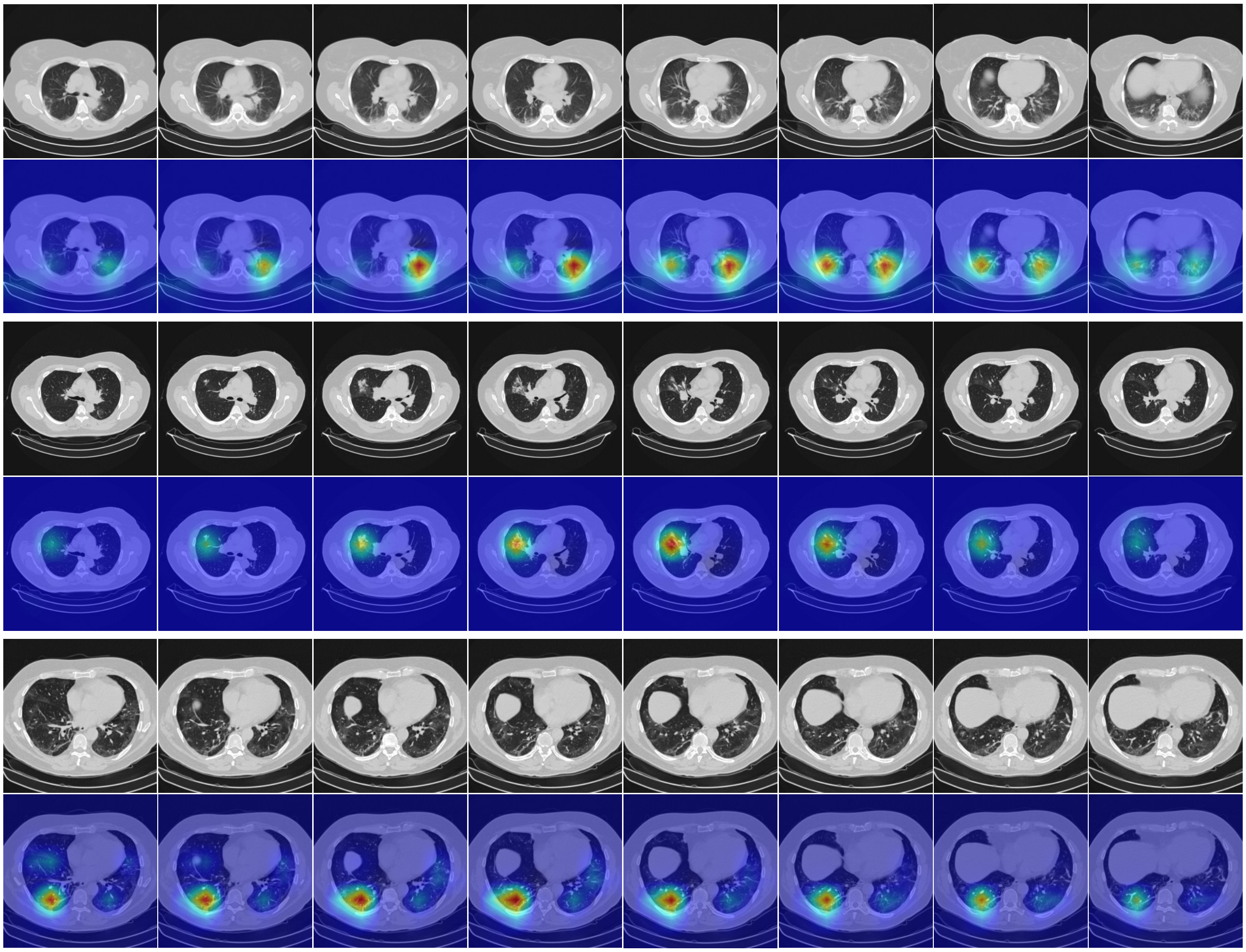

Figure 7. Heatmaps on COVID-19 CT scans.

Processing Workshops (ICASSPW), page 1–5. IEEE, 2023. 4

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2

[4] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021. 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 3

[6] Junlin Hou, Jilan Xu, Rui Feng, Yuejie Zhang, Fei Shan, and Weiya Shi. Cmc-cov19d: Contrastive mixup classification for covid-19 diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 454–461, 2021. 1, 2, 3, 5, 6

[7] Junlin Hou, Jilan Xu, Longquan Jiang, Shanshan Du, Rui Feng, Yuejie Zhang, Fei Shan, and Xiangyang Xue. Periphery-aware covid-19 diagnosis with contrastive representation enhancement. *Pattern Recognition*, 118:108005, 2021. 1

[8] Junlin Hou, Jilan Xu, Nan Zhang, Yi Wang, Yuejie Zhang, Xiaobo Zhang, and Rui Feng. Cmc_v2: Towards more accurate covid-19 detection with discriminative video priors. In *European Conference on Computer Vision*, pages 485–499. Springer, 2022. 2, 5

[9] Junlin Hou, Jilan Xu, Nan Zhang, Yuejie Zhang, Xiaobo Zhang, and Rui Feng. Boosting covid-19 severity detection with infection-aware contrastive mixup classification. In *European Conference on Computer Vision*, pages 537–551. Springer, 2022. 1

[10] Chih-Chung Hsu, Chih-Yu Jian, Chia-Ming Lee, Chi-Han Tsai, and Sheng-Chieh Dai. Strong baseline and bag of tricks for covid-19 detection of ct scans. *arXiv preprint arXiv:2303.08490*, 2023. 2

[11] Menglin Jia, Zuxuan Wu, Austin Reiter, Claire Cardie, Serge Belongie, and Ser-Nam Lim. Exploring visual engagement

signals for representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4206–4217, 2021. 2

[12] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 2

[13] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 4

[14] Dimitrios Kollias, N Bouas, Y Vlaxos, V Brillakis, M Seferis, Ilianna Kollia, Levon Sukissian, James Wingate, and S Kollias. Deep transparent prediction through latent representation analysis. *arXiv preprint arXiv:2009.07044*, 2020. 4

[15] Dimitris Kollias, Y Vlaxos, M Seferis, Ilianna Kollia, Levon Sukissian, James Wingate, and Stefanos D Kollias. Transparent adaptation in deep medical image diagnosis. In *TAILOR*, page 251–267, 2020.

[16] Dimitrios Kollias, Anastasios Arsenos, Levon Soukissian, and Stefanos Kollias. Mia-cov19d: Covid-19 detection through 3-d chest ct image analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, page 537–544, 2021. 1

[17] Dimitrios Kollias, Anastasios Arsenos, and Stefanos Kollias. Ai-mia: Covid-19 detection and severity analysis through medical imaging. In *European Conference on Computer Vision*, page 677–690. Springer, 2022. 1

[18] Dimitrios Kollias, Anastasios Arsenos, and Stefanos Kollias. Ai-enabled analysis of 3-d ct scans for diagnosis of covid-19 & its severity. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, page 1–5. IEEE, 2023. 4

[19] Dimitrios Kollias, Anastasios Arsenos, and Stefanos Kollias. A deep neural architecture for harmonizing 3-d input data analysis and decision making in medical imaging. *Neurocomputing*, 542:126244, 2023. 1, 4

[20] Chun Li, Yunyun Yang, Hui Liang, and Boying Wu. Transfer learning for establishment of recognition of covid-19 on ct imaging using small-sized training datasets. *Knowledge-Based Systems*, 218:106849, 2021. 2

[21] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018. 2

[22] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Efficient parametrization of multi-domain deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8119–8127, 2018. 2

[23] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 6

[24] Ying Song, Shuangjia Zheng, Liang Li, Xiang Zhang, Xiaodong Zhang, Ziwang Huang, Jianwen Chen, Ruixuan Wang, Huiying Zhao, Yutian Chong, et al. Deep learning enables accurate diagnosis of novel coronavirus (covid-19) with ct images. *IEEE/ACM transactions on computational biology and bioinformatics*, 18(6):2775–2780, 2021. 2

[25] Malliga Subramanian, Veerappampalayam Easwaramoorthy Sathishkumar, Jaehyuk Cho, and Kogilavani Shanmugavadivel. Learning without forgetting by leveraging transfer learning for detecting covid-19 infection from ct images. *Scientific Reports*, 13(1):8516, 2023. 2

[26] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 6

[27] Guotai Wang, Jianghao Wu, Xiangde Luo, Xinglong Liu, Kang Li, and Shaoting Zhang. Mis-fm: 3d medical image segmentation using foundation models pretrained on a large-scale unannotated dataset. *arXiv preprint arXiv:2306.16925*, 2023. 2, 3, 5

[28] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 2

[29] Xiaowei Xu, Xiangao Jiang, Chunlian Ma, Peng Du, Xukun Li, Shuangzhi Lv, Liang Yu, Qin Ni, Yanfei Chen, Junwei Su, et al. A deep learning system to screen novel coronavirus disease 2019 pneumonia. *Engineering*, 6(10):1122–1129, 2020. 1, 2

[30] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2736–2746, 2022. 3

[31] Jeffrey O Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. Side-tuning: a baseline for network adaptation via additive side networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 698–714. Springer, 2020. 2

[32] Yunkun Zhang, Jin Gao, Mu Zhou, Xiaosong Wang, Yu Qiao, Shaoting Zhang, and Dequan Wang. Text-guided foundation model adaptation for pathological image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 272–282. Springer, 2023. 2