CMC-COV19D: Contrastive Mixup Classification for COVID-19 Diagnosis

Junlin Hou[1#], Jilan Xu[1#], Rui Feng[1*], Yuejie Zhang[1*], Fei Shan[2], Weiya Shi[2]
[1]School of Computer Science, Shanghai Key Lab of Intelligent Information Processing, Fudan University
[2]Department of Radiology, Shanghai Public Health Clinical Center, Fudan University
{jlhou18,jilanxu18,fengrui,yjzhang}@fudan.edu.cn,shanfei_2901@163.com,shiweiya@shphc.org.cn

Abstract

Deep learning methods have been extensively investigated for rapid and precise computer-aided diagnosis during the outbreak of the COVID-19 epidemic. However, there are still remaining issues to be addressed, such as distinguishing COVID-19 in the complex scenario of multi-type pneumonia classification. In this paper, we aim to boost the COVID-19 diagnostic performance with more discriminative deep representations of COVID and non-COVID categories. We propose a novel COVID-19 diagnosis approach with contrastive representation learning to effectively capture the intra-class similarity and inter-class difference. Besides, we design an adaptive joint training strategy to integrate the classification loss, mixup loss, and contrastive loss. Through the joint loss function, we obtain the high-level representations which are highly discriminative in COVID-19 screening. Extensive experiments on two chest CT image datasets, i.e., CC-CCII dataset and COV19-CT-DB database, demonstrate the effectiveness of our proposed approach in COVID-19 diagnosis. ***Our method won the first prize in the ICCV 2021 Covid-19 Diagnosis Competition of AI-enabled Medical Image Analysis Workshop****. Our code is publicly available at* `https://github.com/houjunlin/Team-FDVTS-COVID-Solution`.

1. Introduction

The Coronavirus Disease 2019 SARS-CoV-2 (COVID-19) has become a global pandemic with an exponential growth and mortality rate. Early detection and treatment are of great importance to the slowdown of viral transmission and the control of the disease. As a reliable complement to the Reverse Transcription-Polymerase Chain Reaction (RT-PCR) testing, thoracic computed tomography (CT) has also been recognized to be a powerful tool for clinical diagnosis, especially in many hard-hit regions. CT scans can provide a detailed overview of the internal structure of lung parenchyma. Some characteristic radiological patterns of COVID-19 can be clearly detected in chest CT images, including ground glass opacities in the lung periphery, rounded opacities, enlarged intra-infiltrate vessels, and later more consolidations. However, the large volume of CT scan images requires a heavy workload on radiologists and physicians to diagnose COVID-19. Meanwhile, with the rapid increase in the number of new and suspected COVID-19 cases, it is evident that there is an urgent need for detecting COVID-19 from CT scans automatically using deep learning approaches.

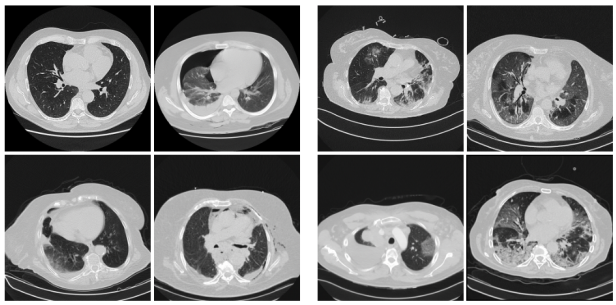

Figure 1. Samples of non-COVID and COVID from the COV19-CT-DB database.

Deep learning approaches have demonstrated significant improvement in fighting against COVID-19. They have been widely applied to the lung and infection region segmentation, the clinical diagnosis and assessment, as well as the pioneering basic and clinical research [19]. For instance, Javaheri *et al.* [9] designed the CovidCTNet to differentiate COVID-19 from community-acquired pneumonia and other lung diseases. Li *et al.* [17] proposed a ResNet50-based COVNet, which extracted both 2D local features and 3D global features to identify COVID-19 from CAP and non-pneumonia. Xu *et al.* [25] developed a deep learning system to categorize COVID-19, Influenza-A viral pneumonia, and healthy cases with a location-attention mechanism. Besides, Qian *et al.* [18] proposed a Multi-task

*Corresponding authors. # Equal contribution.

Multi-slice Deep Learning System (M3Lung-Sys), which is composed of slice-level and patient-level networks for multi-class pneumonia screening. However, as a new disease, COVID-19 still has some similar manifestations with various types of pneumonia, as illustrated in Fig. 1. It increases the challenge of accurate COVID-19 diagnostic performance. Thus, it is important to learn more discriminative representations of COVID-19 and other pneumonia.

This paper summarizes a novel solution to improve COVID-19 detection performance, based on the recently proposed COVID-19 diagnosis approach [8]. We introduce contrastive representation learning to discover similar patterns in COVID-19 cases and differences from other categories of pneumonia. Besides, we propose an adaptive joint training strategy that combines the classification loss, mixup loss, and contrastive loss to discriminate features of COVID-19 volumes from those of the non-COVID cases. We further employ an inflated 3D ImageNet pre-trained ResNest50 [27] as a strong feature extractor to boost more accurate COVID-19 diagnostic performance. By combing these techniques, our method achieved superior diagnostic results and won the first prize in the ICCV 2021 Covid-19 Diagnosis Competition of AI-enabled Medical Image Analysis Workshop.

The remainder of this paper is organized as follows. Section 2 reviews some related works. Section 3 describes our proposed method with the contrastive representation learning and the adaptive joint training strategy in detail. Section 4 presents the two databases used in this work. Section 5 provides implementation settings and experimental results of COVID-19 detection. Section 6 concludes our work.

2. Related Work

We first briefly review some representative works using deep learning methods to diagnose COVID-19 in CT images. Then, we present a literature review on the highly relevant concepts from the perspective of methodology.

2.1. COVID-19 Screening

During the outbreak of COVID-19, numerous deep learning approaches have been intensively and rapidly conducted on COVID-19 screening. Generally, the existing methods for COVID-19 diagnosis with CT images can be roughly divided into two categories. One category is to segment the infection regions first, and then utilize them for the COVID-19 classification. For example, Zhang *et al.* [28] and Jin *et al.* [10] both developed the combined "segmentation-classification" model pipelines to diagnose COVID-19. Chen *et al.* [2] detected the suspicious lesions on three consecutive CT scans to determine COVID-19 by training a UNet++ model. This category of work requires a large number of segmentation annotations. However, it is extremely time-consuming and difficult to acquire accurate segmentation annotations.

The second category is to train the classification model to diagnose COVID-19 directly, using 3D volume-level or 2D slice-level supervision. Representatively, Wang *et al.* [21] built a modified Inception model to distinguish COVID-19 from typical viral pneumonia using transfer learning technique. Song *et al.* [20] proposed a diagnosis system named DeepPneumonia to detect COVID-19 patients from bacterial pneumonia and healthy controls. In addition, Wang *et al.* [22] used a 3D DeCoVNet to detect COVID-19, which took a CT volume with its lung mask as input. In this work, we follow the second category to diagnose COVID-19 with only volume-level or slice-level supervision, as it is expensive to annotate labels in a voxel-wise or pixel-wise manner.

2.2. Contrastive Learning

Contrastive learning aims to learn representations by contrasting positive and negative sample pairs. It is at the core of recent works on the instance-level discrimination of self-supervised learning. Instance-level discrimination aims to learn representations by imposing transformation invariances in the latent space. The augmented images from the same image are regarded as the positives, which should be closer together than those from different images in the representation space. The Exemplar CNN [5] represented each instance as a vector and trained a network to recognize each instance. Since negative sampling is an important issue in contrastive learning, Wu *et al.* [24] constructed a memory bank to store instance vectors. He *et al.* [6] proposed the momentum contrast (MoCo) for visual representational learning. Recently, Chen *et al.* [3] proposed the SimCLR, which generated augmentation-invariant embeddings for input images. These works achieved great empirical successes in self-supervised representation learning.

The potential of contrastive learning on COVID-19 screening has also been explored in some research works. For example, He *et al.* [7] integrated contrastive learning and transfer learning to pre-train the classification networks for COVID-19 diagnosis. Wang *et al.* [23] used contrastive learning to tackle the cross-site domain difference when diagnosing COVID-19 on heterogeneous CT datasets. Chen *et al.* [4] adopted a prototypical network pre-trained by the momentum contrastive learning method [6] for few-shot COVID-19 diagnosis. Li *et al.* [16] presented a contrastive multi-task CNN which can improve the generalization on unseen CT or X-ray samples to diagnose COVID-19.

In our work, we fully exploit contrastive learning to obtain more discriminative representations in chest CT images of non-COVID and COVID types. In addition to the general contrastive learning, we further introduce the category information for better exploration of intra-class similarity and inter-class difference. Particularly, we adopt it as an

auxiliary learning task that can effectively improve the classification performance of COVID-19.

3. Methodology

3.1. Overview

To make a more accurate diagnosis of COVID-19, we construct a novel deep learning network, namely CMC-COV19D, with contrastive representation learning (CRL) and mixup pneumonia classification, as illustrated in Figure 2. The CMC-COV19D network takes the 3D CT volume as input and outputs the diagnosis results of non-COVID and COVID. The encoder of CMC-COV19D is initialized by the inflated 3D ImageNet pre-trained weights. Our network is trained by an adaptive joint loss which is composed of the classification loss, mixup loss, and contrastive loss.

3.2. Contrastive Representation Learning

To learn more discriminative representations of non-COVID and COVID, we develop the contrastive representation learning (CRL) as an auxiliary task for a precise COVID-19 diagnosis. Our novel CMC-COV19D with CRL is comprised of the following components.

- A stochastic *data augmentation* module, $A(\cdot)$, which transforms an input CT sample x into a randomly augmented sample $\tilde{x}$. We generate two augmented volumes from each input CT sample. Specifically, we sequentially apply three augmentations for CT samples: 1) random cropping on the transverse plane followed by resizing back to the original resolution; 2) random cropping on the vertical section to a fixed depth d; and 3) random changes in brightness and contrast.
- A neural network *base encoder*, $E(\cdot)$, which maps the augmented CT sample $\tilde{x}$ to a representation vector $r = E(\tilde{x}) \in \mathbb{R}^{d_e}$ in the d_e-dimensional latent space. The augmented CT samples of different categories share the same encoder and generate pairs of representation vectors. In principle, any CNN architecture can be adopted as the encoder here, of which the outputs of the average pooling layer are used as the representation vectors. Then, the representations are normalized to the unit hypersphere.
- A *projection network*, $P(\cdot)$, which is used to map the representation vector r to a relative low-dimension vector $z = P(r) \in \mathbb{R}^{d_p}$ for the contrastive loss computation. A multi-layer perception (MLP) can be employed as the projection network. This vector is also normalized to the unit hypersphere, which enables the inner product to measure distances in the projection space. This subnetwork is only used for the contrastive loss.
- A *classifier*, $C(\cdot)$, which classifies the representation vector $r \in \mathbb{R}^{d_e}$ to the pneumonia prediction and mixup prediction. It is composed of the fully connected layer and the Softmax operation.

Given a minibatch of N randomly sampled CT images and their pneumonia-type labels $\{(x_k, y_k)\}_{k=1,\ldots,N}$, we can generate a minibatch of $2N$ samples $\{(\tilde{x}_{2k-1}, \tilde{y}_{2k-1}), (\tilde{x}_{2k}, \tilde{y}_{2k})\}_{k=1,\ldots,N}$ after performing data augmentations, where $\tilde{x}_{2k-1}$ and $\tilde{x}_{2k}$ are two random augmented CT samples of x_k, and $\tilde{y}_{2k-1} = \tilde{y}_{2k} = y_k$.

In this work, our goal is to enhance the representation similarity within the same category, and simultaneously increase the inter-class difference between different types. Therefore, we introduce the label information into contrastive learning. We redefine the positives as any two augmented CT samples from the same category, whereas the CT samples from different classes are considered as negative pairs. Assume that two samples x_i and x_j from the same class are considered as a positive pair, the contrastive loss is optimized when x_i is similar to its positive sample x_j and dissimilar to the other negative samples. Let $i \in \{1, \ldots, 2N\}$ be the index of an arbitrary augmented sample, the contrastive loss function is defined as:

$$\mathcal{L}_{con} = \frac{1}{2N} \sum_{i=1}^{2N} \mathcal{L}_{con}^{i}, \tag{1}$$

$$\mathcal{L}_{con}^{i} = \frac{-1}{2N_{\tilde{y}_i} - 1} \sum_{j=1}^{2N} \mathbb{1}_{i \neq j} \cdot \mathbb{1}_{\tilde{y}_i = \tilde{y}_j} \cdot \log \frac{\exp(z_i^T \cdot z_j / \tau)}{\sum_{k=1}^{2N} \mathbb{1}_{i \neq k} \cdot \exp(z_i^T \cdot z_k / \tau)}, \tag{2}$$

where $z = P(E(\tilde{x}))$ is the representation vector; $\mathbb{1} \in \{0, 1\}$ is an indicator function, and $\tau > 0$ denotes a scalar temperature hyper-parameter. $N_{\tilde{y}_i}$ is the total number of samples in a minibatch that have the same label $\tilde{y}_i$. In this way, samples z_i and z_j have the same label (i.e., $\tilde{y}_i = \tilde{y}_j$), and they are considered as the positive pair. We apply the inner product to measure the similarity between the normalized vectors z_i and z_j in d_p-dimensional space. The final contrastive loss is calculated as the summation of the loss over all pairs of indices (i, j) and (j, i) after the Softmax function. Within the context of the contrastive loss, the encoder is trained to maximize the similarity between positive samples, while minimizing the similarity between negative pairs simultaneously.

3.3. Adaptive Joint Training Strategy

To improve the pneumonia classification accuracy, we particularly design an adaptive joint training strategy. Besides the aforementioned contrastive loss, we first introduce each individual loss.

Classification loss The classifier C is learned to predict the classification results $\hat{y}$ using the standard cross-entropy

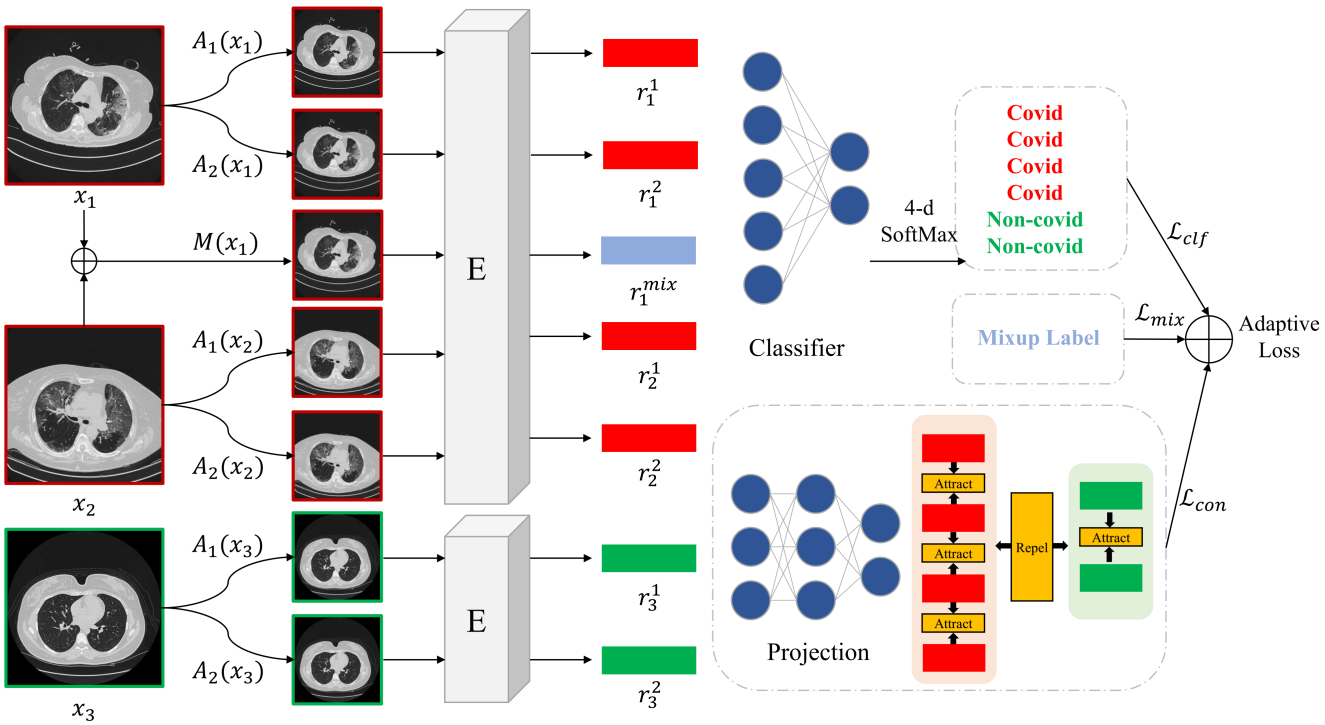

Figure 2. An overview of our CMC-COV19D network architecture (best viewed in colour). Firstly, the CT image is transformed into two volumes. Each augmented CT image is then fed into the ImageNet pre-trained encoder to generate corresponding high-level representations. The mixup technique is adopted to obtain the mixup representations. A classifier is learned on top of the representations for pneumonia classification, resulting in $\mathcal{L}_{clf}$ and $\mathcal{L}_{mix}$. Meanwhile, the representations are mapped by the projection network into the d_p-dimensional space to calculate the contrastive loss $\mathcal{L}_{con}$.

loss $\mathcal{L}_{ce}$, which is defined as:

$$\mathcal{L}_{clf} = \frac{1}{2N}\sum_{i=1}^{2N}\mathcal{L}_{ce}^{i}, \tag{3}$$

$$\mathcal{L}_{ce}^{i} = -\tilde{y}_i^T \log \hat{y}_i, \tag{4}$$

where $\tilde{y}_i$ denotes the one-hot vector of ground truth label, and $\hat{y}_i$ is predicted probability of the sample x_i $(i = 1, \ldots, 2N)$.

Mixup loss To further boost the generalization ability of the model, we adopt the mixup [26] strategy during training. As defined in Section 3.2, for randomly augmented CT samples $\tilde{x}_{2k-1}$ and $\tilde{x}_{2k}$, we generate the pseudo mixup samples and their labels as:

$$\begin{aligned}
\tilde{x}_{2k-1}^{mix} &= \lambda\tilde{x}_{2k-1} + (1-\lambda)\tilde{x}_p, \\
\tilde{y}_{2k-1}^{mix} &= \lambda\tilde{y}_{2k-1} + (1-\lambda)\tilde{y}_p, \\
\tilde{x}_{2k}^{mix} &= \lambda\tilde{x}_{2k} + (1-\lambda)\tilde{x}_q, \\
\tilde{y}_{2k}^{mix} &= \lambda\tilde{y}_{2k} + (1-\lambda)\tilde{y}_q,
\end{aligned} \tag{5}$$

where p and q are randomly selected indices. The mixup loss is then defined as the average of the standard cross entropy losses of the mixup samples:

$$\mathcal{L}_{mix} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\mathcal{L}_{ce}(\tilde{x}_{2i-1}^{mix}, \tilde{y}_{2i-1}^{mix}) + \frac{1}{2}\mathcal{L}_{ce}(\tilde{x}_{2i}^{mix}, \tilde{y}_{2i}^{mix}). \tag{6}$$

Different from the original design [26] where they replaced the classification loss with the mixup loss, we merge the mixup loss with the classfication loss to enhance the classification ability on both raw samples and mixup samples.

Adaptive loss The network is trained by three loss functions, i.e., $\mathcal{L}_{con}$ for the representation learning, $\mathcal{L}_{clf}$ for the pneumonia classification, and $\mathcal{L}_{mix}$ for the mixup prediction. To merge CRL loss, classification loss, and mixup loss effectively, we design a combined objective function with adaptive weights [11] to balance the three loss functions for better joint learning performance:

$$\mathcal{L}_{joint} = \frac{1}{\sigma_1^2}\mathcal{L}_{con} + \frac{1}{\sigma_2^2}(\mathcal{L}_{clf} + \mathcal{L}_{mix}) + \log\sigma_1 + \log\sigma_2, \tag{7}$$

where σ_1 and σ_2 are utilized to learn the relative weights of the three losses adaptively. The adaptive joint training strategy relieves the manual effort of tuning the balanced weights.

3.4. Techniques to Boost Diagnostic Performance

In this subsection, we briefly introduce two key techniques used in the ICCV 2021 COV19D challenge to further boost the diagnostic performance of our model.

3.4.1 Inflated Weight Transfer

Transfer learning is a commonly adopted method to tackle insufficient data. For example, ImageNet pre-trained models have been applied successfully in image classification, object detection, and semantic segmentation. Here, we adopt ImageNet pre-trained 2D ResNest [27], a recently proposed variant of ResNet, as the robust feature encoder. With the goal of adapting pre-trained 2D weights to 3D, we inflate the 2D weights along the voxel dimension for common convolution layers [1]. In terms of the split-attention blocks [27], we maintain the cardinal groups and radix during the inflation. To this end, the inflated 3D ResNest serves as a strong feature encoder.

3.4.2 Intra-model and Inter-model Ensemble

Aiming to boost the generalization ability of the model on the testing set, we apply the intra-model and inter-model ensemble strategies. As our proposed contrastive representation learning encourages CT volumes of the same category to have similar representations despite different augmentations of each volume. Thus, we also apply different augmentations $\{A_1(x), A_2(x), ..., A_t(x)\}$ to each volume. We adopt $t = 3$ in our experiment. In spite of different appearances, these augmented samples are semantically related. We term the intra-model ensemble as the average of the model's predictions on these samples. Then, we perform inter-model ensemble by aggregating the predicted results of multiple models. In the challenge, we simply adopt the average operation as the default aggregation strategy.

4. Datasets

In this work, we use two datasets, the publicly available CC-CCII dataset [28] and COV19-CT-DB dataset in "AI-enabled Medical Image Analysis Workshop and Covid-19 Diagnosis Competition (MIA-COV19D)" [12].

4.1. COV19-CT-DB Database

COVID19-CT-Database (COV19-CT-DB) consists of chest CT scans marking the existence of COVID-19. Data collection was conducted in the period from September 1, 2020 to March 31, 2021. The COV19-CT-DB database consists of about 5,000 chest CT scan series from over 1,000 patients and 2,000 subjects. Each CT slice was annotated by 4 experienced medical experts. The experts showed a high degree of agreement (around 98%) on the annotations. The number of slices of each 3D scan ranges from 50 to 700. The training set contains 1,560 3D CT scans (690 COVID-19 cases and 870 Non-COVID-19 cases). The validation set consists of 374 3D CT scans (165 COVID-19 cases and 209 Non-COVID-19 cases). The testing set includes 3,455 scans and the labels are not available during the challenge.

4.2. CC-CCII Dataset

CC-CCII dataset [28] contains 444,034 slices from 4,356 CT scans of 2,778 people, including 1,578 COVID-19 scans, 1,614 common pneumonia scans, and 1,164 normal control scans. Patients were randomly divided into the training set (80%), the validation set (10%), and the testing set (10%). Among all the CT slices, a section of 750 CT slices from 150 COVID-19 patients was manually segmented into the background, lung field, ground-glass opacity (GGO), and consolidation (CL). All images are stored in the PNG, JPG, and TIFF file formats.

5. Experiments

5.1. Implementation Details

Our data pre-processing procedure is as follows. First, each 2D chest CT scan series is composed into a 3D volume of shape (D, H, W), where D, H, W denotes the slice, height, and width, respectively. Then, each volume is resized from its original size to $(128, 256, 256)$. Finally, we transform the CT volume to the interval $[0, 1]$ for intensity normalization. We apply inflated 3D ResNet50 and ResNest50 as the backbones in our experiments. The value of parameter d_e is 2,048 and d_p is 128. We optimize the network using the Adam algorithm with a weight decay of 10^{-5}. The network is trained for 100 epochs. The initial learning rate is set to 0.001 and then divided by 10 at 30% and 80% of the total number of training epochs. We adopt the accuracy and Macro F1 as the evaluation metrics. The Macro F1 is defined as the unweighted average of the class-wise/label-wise F1 Scores, i.e., the unweighted average of the COVID-19/non-COVID-19 class F1 Score. Our methods are implemented in PyTorch and run on four NVIDIA Tesla V100 GPUs.

5.2. Results on COV19-CT-DB

Table 1 shows the results of our methods and the baseline model on the validation set of COV19-CT-DB database. The baseline approach is based on a CNN-RNN architecture that performs 3D CT scan analysis. The method follows the work [13, 15, 14] on developing deep neural architectures for predicting COVID-19. It can be seen that our proposed model with ResNet50 backbone achieves 0.91 on Macro F1, which obtains the significant improvement of 21% compared with the baseline ResNet50-GRU model.

In order to explore the effect of backbone, we apply the recently proposed ImageNet pre-trained Split-

Methods	Backbone	Depth	$\mathcal{L}_{clf}$	$\mathcal{L}_{con}$	$\mathcal{L}_{mix}$	Accuracy	Macro F1
Baseline	ResNet50+GRU	-	$\surd$	$\times$	$\times$	-	70.00
Model1	ResNet50	128	$\surd$	$\surd$	$\times$	91.71 [89.30, 94.12]	91.53 [89.04, 94.03]
Model2	ResNest50	64	$\surd$	$\times$	$\surd$	91.18 [88.50, 93.58]	91.06 [88.39, 93.38]
Model3	ResNest50	64	$\surd$	$\surd$	$\times$	92.25 [89.84, 94.65]	92.10 [89.66, 94.51]
Model4	ResNest50	64	$\surd$	$\surd$	$\surd$	94.65 [92.78, 96.52]	94.54 [92.48, 96.42]
Model5	Ensemble	-	-	-	-	93.05 [90.91, 95.19]	92.88 [90.56, 95.03]

Table 1. The comparison results on the validation set of COV19-CT-DB database. The values in brackets are 95% confidence intervals.

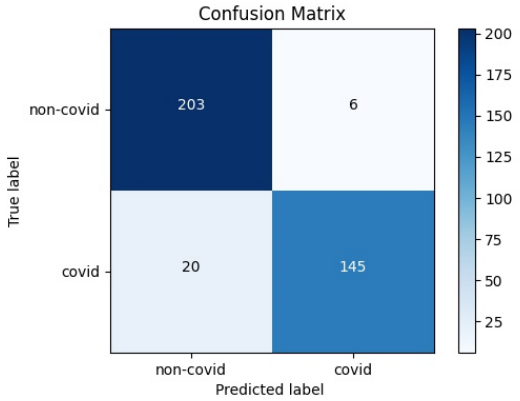

Figure 3. The confusion matrix of model's prediction.

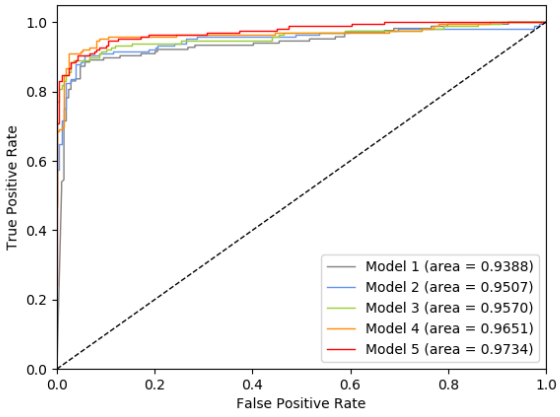

Figure 4. The roc curve of model's prediction.

Attention Network [27], a.k.a. ResNest50, to our approach. ResNest50 achieves better performance compared with ResNet-based networks with 0.5% and 0.6% improvement on accuracy and Macro F1, respectively. Surprisingly, by introducing the mixup joint loss, Model4 obtains the state-of-the-art results with 94.65% on accuracy and 94.54% on Macro F1. Finally, we perform intra-model and inter-model ensemble using Model1, Model3, and Model4 to obtain Model5, as illustrated in the last row in Table 1. The confusion matrix and roc curve on the validation set

Rank	Teams	Macro F1	F1 (C)	F1 (NC)
1	**FDVTS_COVID**	**90.43**	**83.60**	**97.27**
2	SenticLab.UAIC	90.06	82.96	97.17
3	ACVLab	88.74	80.63	96.84
4	DeepCam	88.22	79.79	96.64
5	TAC	87.77	78.78	96.75
6	LoVE	84.20	73.65	94.76
7	Heal it	78.86	63.65	94.06
8	HCMUS-HGV	78.13	63.08	93.18
9	Blessing	75.67	58.95	92.40
10	AvengerQ	71.83	51.63	92.04
11	Terps	70.86	48.96	92.75
12	x Vision	70.50	53.67	87.33
13	baseline	67.00	54.38	79.62

Table 2. The competition results on the testing set of the COV19-CT-DB database. The last two columns show the F1 Score on COVID and non-COVID cases, respectively.

is shown in Fig 3 and Fig 4, respectively. Our proposed Model5 successfully classifies most non-covid (203) and covid (145) cases with an average precision of 0.931. However, as shown in Figure 3, our approach makes 6 false-positive predictions. We recognize that some false-positive cases belong to other types of pneumonia, which makes it challenging for the model to distinguish them from covid-19 cases. Also, Model5 achieved the highest AUC among the models, as illustrated in Fig 4.

It can be seen from table 2 that our proposed Model5 ranked the first in the challenge, surpassing the second team by 0.37 on the Macro F1. Compared to other methods, our model achieves significant improvement on the Macro F1 (COVID), indicating the ability to distinguish COVID cases from other types of pneumonia correctly. Figure 5 shows the visualization of the classification results using Gradient-weighted Class Activation Mapping (Grad-CAM). One can see that our model shows a great localization ability of suspicious lesions. It manages to localize the lesions that usually appear at the periphery of the lung. This suggests the clear interpretability of the diagnosis results from our proposed model. The attention maps can be possibly used as the basis to derive the COVID-19 diagnosis in clinical practice.

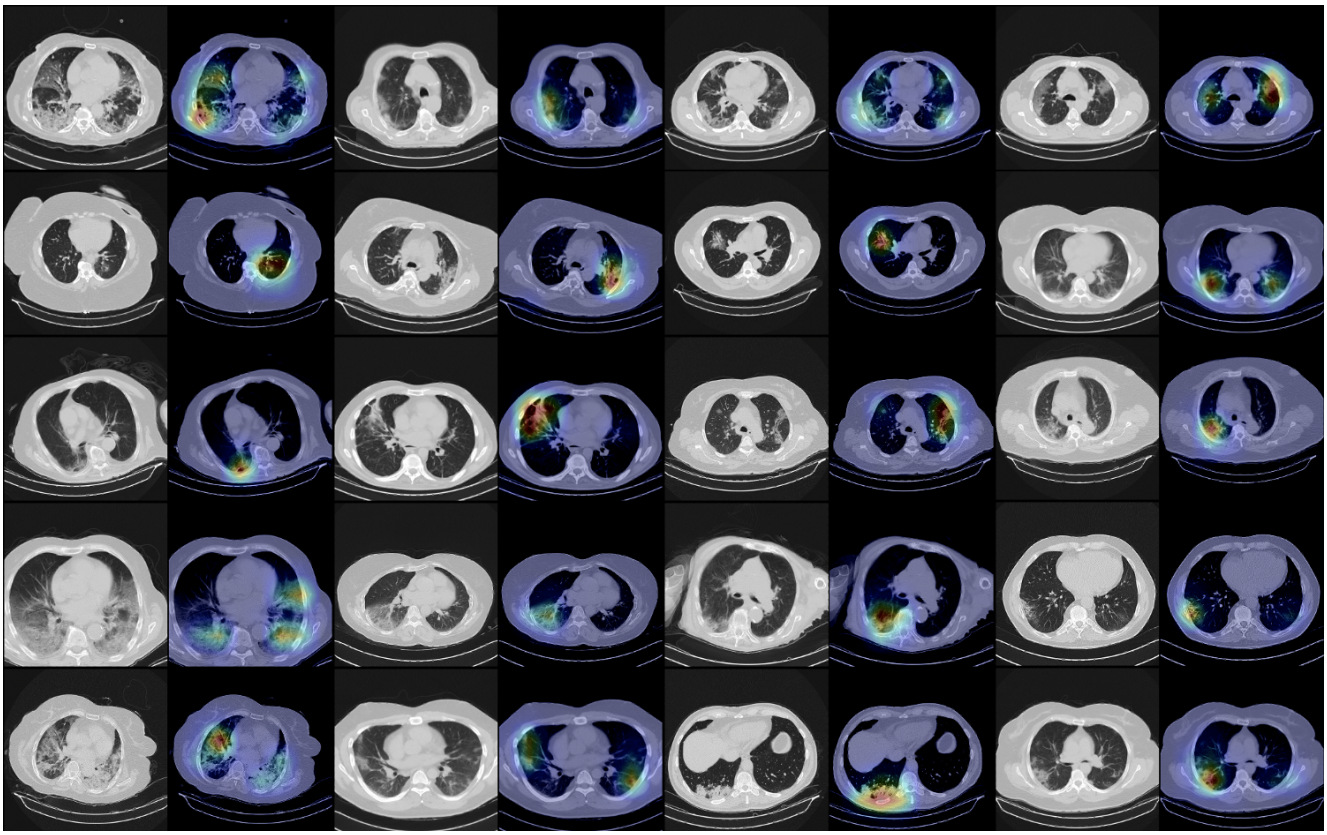

Figure 5. The visualization results on COVID-19 cases.

Methods	Acc	Sen	Spec	AUC
Zhang *et al.* [28]	92.49	94.93	91.13	97.97
Li *et al.* [16]	93.46	90.57	90.84	89.22
CMC-COV19D	**96.85**	**96.47**	**97.08**	**98.83**

Table 3. The comparison results on the CC-CCII database.

5.3. Results on CC-CCII

Table 3 shows the results of our Model4 and other existing approaches on the CC-CCII dataset. We conduct the binary diagnosis which denotes the classification between non-COVID-19 and COVID-19. The category of non-COVID-19 consists of common pneumonia and healthy cases. Our CMC-COV19D achieves 96.85% Acc, 96.47% Sen, 97.08% Spec, and 98.83% AUC, which surpasses the other approaches by a large margin. The binary classification task demonstrates the superiority of our method on the CC-CCII dataset.

6. Conclusion

In this work, we propose a novel COVID-19 diagnosis approach named CMC-COV19D. Particularly, we design contrastive representation learning (CRL) as an auxiliary task and propose an adaptive joint training strategy. CRL extends the general contrastive learning, which can capture the intra-class similarity and inter-class difference. CRL is also generalizable to multi-type pneumonia diagnosis. The adaptive joint training strategy integrates an adaptive combination of the classification loss, mixup loss, and contrastive loss. Extensive experiments on the COVID19-CT-DB database and CC-CCII dataset show the effectiveness of our model in distinguishing COVID-19 cases.

Acknowledgement

This work was supported by National Natural Science Foundation of China (No.61976057), the Science and Technology Commission of Shanghai Municipality (No.21511101000, No.20511100800, No.20511101203, No.19DZ2205700), the Science and Technology Major Project of Commission of Science and Technology of Shanghai (No.2021SHZDZX0103). Rui Feng and Yuejie Zhang are corresponding authors.

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *pro-*

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6299–6308, 2017.

[2] Jun Chen, Lianlian Wu, Jun Zhang, Liang Zhang, Dexin Gong, Yilin Zhao, et al. Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography. *Scientific Reports*, 10(1):1–11, 2020.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607, 2020.

[4] Xiaocong Chen, Lina Yao, Tao Zhou, Jinming Dong, and Yu Zhang. Momentum contrastive learning for few-shot covid-19 diagnosis from chest ct images. *Pattern Recognition*, 113:107826, 2021.

[5] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox. Discriminative unsupervised feature learning with exemplar convolutional neural networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 38(9):1734–1747, 2014.

[6] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE Conference on Computer Vision and Pattern Recognition*, pages 9726–9735, 2020.

[7] Xuehai He, Xingyi Yang, Shanghang Zhang, Jinyu Zhao, Yichen Zhang, Eric Xing, et al. Sample-efficient deep learning for covid-19 diagnosis based on ct scans. *MedRxiv*, 2020.

[8] 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.

[9] T. Javaheri, M. Homayounfar, Z. Amoozgar, R. Reiazi, and R. Rawassizadeh. Covidctnet: An open-source deep learning approach to identify covid-19 using ct image. 2020.

[10] Shuo Jin, Bo Wang, Haibo Xu, Chuan Luo, Lai Wei, Wei Zhao, et al. Ai-assisted ct imaging analysis for covid-19 screening: building and deploying a medical ai system in four weeks. *MedRxiv*, 2020.

[11] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018.

[12] Dimitrios Kollias, Anastasios Arsenos, Levon Soukissian, and Stefanos Kollias. Mia-cov19d: Covid-19 detection through 3-d chest ct image analysis. *arXiv preprint arXiv:2106.07524*, 2021.

[13] 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.

[14] Dimitrios Kollias, Athanasios Tagaris, Andreas Stafylopatis, Stefanos Kollias, and Georgios Tagaris. Deep neural architectures for prediction in healthcare. *Complex & Intelligent Systems*, 4(2):119–131, 2018.

[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*, pages 251–267, 2020.

[16] Jinpeng Li, Gangming Zhao, Yaling Tao, Penghua Zhai, Hao Chen, Huiguang He, et al. Multi-task contrastive learning for automatic ct and x-ray diagnosis of covid-19. *Pattern Recognition*, 114:107848, 2021.

[17] Lin Li, Lixin Qin, Zeguo Xu, Youbing Yin, Xin Wang, Bin Kong, et al. Artificial intelligence distinguishes covid-19 from community acquired pneumonia on chest ct. *Radiology*, 296:200905, 2020.

[18] Xuelin Qian, Huazhu Fu, Weiya Shi, Tao Chen, Yanwei Fu, Fei Shan, et al. M^3lung-sys: A deep learning system for multi-class lung pneumonia screening from ct imaging. *IEEE Journal of Biomedical and Health Informatics*, 24(12):3539–3550, 2020.

[19] Feng Shi, Jun Wang, Jun Shi, Ziyan Wu, Qian Wang, Zhenyu Tang, et al. Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for covid-19. *IEEE Reviews in Biomedical Engineering*, pages 1–1, 2020.

[20] Ying Song, Shuangjia Zheng, Liang Li, Xiang Zhang, Xiaodong Zhang, Ziwang Huang, et al. Deep learning enables accurate diagnosis of novel coronavirus (covid-19) with ct images. *MedRxiv*, 2020.

[21] Shuai Wang, Bo Kang, Jinlu Ma, Xianjun Zeng, Mingming Xiao, Jia Guo, et al. A deep learning algorithm using ct images to screen for corona virus disease (covid-19). *European radiology*, pages 1–9, 2021.

[22] X. Wang, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, et al. A weakly-supervised framework for covid-19 classification and lesion localization from chest ct. *IEEE Transactions on Medical Imaging*, 39(8):2615–2625, 2020.

[23] Zhao Wang, Quande Liu, and Qi Dou. Contrastive cross-site learning with redesigned net for covid-19 ct classification. *IEEE Journal of Biomedical and Health Informatics*, 24(10):2806–2813, 2020.

[24] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[25] Xiaowei Xu, Xiangao Jiang, Chunlian Ma, Peng Du, Xukun Li, Shuangzhi Lv, et al. A deep learning system to screen novel coronavirus disease 2019 pneumonia. *Engineering*, 6(10):1122–1129, 2020.

[26] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[27] 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. *arXiv preprint arXiv:2004.08955*, 2020.

[28] Kang Zhang, Xiaohong Liu, Jun Shen, Zhihuan Li, Ye Sang, Xingwang Wu, et al. Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography. *Cell*, 181(6):1423–1433, 2020.