Efficient and Effective Training of COVID-19 Classification Networks With Self-Supervised Dual-Track Learning to Rank

Yuexiang Li, Dong Wei, Jiawei Chen, Shilei Cao, Hongyu Zhou, Yanchun Zhu, Jianrong Wu, Lan Lan, Wenbo Sun, Tianyi Qian, Kai Ma, Haibo Xu, and Yefeng Zheng

***Abstract*—Coronavirus Disease 2019 (COVID-19) has rapidly spread worldwide since first reported. Timely diagnosis of COVID-19 is crucial both for disease control and patient care. Non-contrast thoracic computed tomography (CT) has been identified as an effective tool for the diagnosis, yet the disease outbreak has placed tremendous pressure on radiologists for reading the exams and may potentially lead to fatigue-related mis-diagnosis. Reliable automatic classification algorithms can be really helpful; however, they usually require a considerable number of COVID-19 cases for training, which is difficult to acquire in a timely manner. Meanwhile, how to effectively utilize the existing archive of non-COVID-19 data (the negative samples) in the presence of severe class imbalance is another challenge. In addition, the sudden disease outbreak necessitates fast algorithm development. In this work, we propose a novel approach for effective and efficient training of COVID-19 classification networks using a small number of COVID-19 CT exams and an archive of negative samples. Concretely, a novel self-supervised learning method is proposed to extract features from the COVID-19 and negative samples. Then, two kinds of soft-labels ('difficulty' and 'diversity') are generated for the negative samples by computing the earth mover's distances between the features of the negative and COVID-19 samples, from which data 'values' of the negative samples can be assessed. A pre-set number of negative samples are selected accordingly and fed to the neural network for training. Experimental results show that our approach can achieve superior performance using about half of the negative samples, substantially reducing model training time.**

***Index Terms*—Coronavirus disease 2019, Self-supervised learning, efficient network training.**

I. Introduction

CORONAVIRUS Disease 2019 (COVID-19) has rapidly spread all over the world since first reported in December 2019. The disease is highly contagious and has infected over 14.4 million people worldwide to date (Jul 20th, 2020).[1] On March 11th, 2020, the World Health Organization (WHO) declared COVID-19 a pandemic. The reverse transcription polymerase chain reaction (RT-PCR) assay of sputum or nasopharyngeal swab is considered as the gold standard for confirmation of COVID-19 cases. However, several studies have reported insufficient sensitivities (i.e., high false negative rates) of RT-PCR for effective early diagnosis and subsequent treatment of presumptive patients [1]–[3]. Meanwhile, as most COVID-19 patients exhibit respiratory symptoms (mainly pneumonia) [4], the non-contrast thoracic computed tomography (CT) becomes an alternative solution to help diagnose COVID-19 positive patients from suspected cohorts.

Several studies [2], [5] described typical chest CT findings of COVID-19 as diffuse or focal ground-glass opacities, particularly bilateral and peripheral ground-glass, as well as consolidative pulmonary opacities. In addition, Bernheim *et al.* [5] noted the progression of chest CT manifestations from early to late stages, characterized by greater lung involvement, crazing paving and reverse halo signs, etc. Moreover, both Xie *et al.* [1] and Fang *et al.* [2] reported superior sensitivities of non-contrast chest CT to RT-PCR. However, due to the sudden outbreak of the COVID-19 pandemic, radiologists are now facing great pressure in reading the tremendous quantity of CT exams—a thin-slice CT sequence of 300 slices can take 5–15 minutes of a radiologist to make the diagnosis. Besides being time-consuming and painstaking, the reading process can become error- and omission-prone considering the extremely

Manuscript received May 22, 2020; revised July 20, 2020; accepted August 15, 2020. Date of publication August 20, 2020; date of current version October 5, 2020. This work was supported in part by the Key Area Research and Development Program of Guangdong Province, China under Grant 2018B010111001, in part by National Key Research and Development Project under Grant 2018YFC2000702, in part by the Natural Science Foundation of China under Grant 61702339, in part by the Science and Technology Program of Shenzhen, China under Grant ZDSYS201802021814180, and in part by Emergency Science and Technology Project on COVID-19 Pneumonia supported by the Department of Science and Technology of Hubei Province Project under Grant 2020FCA016. *(Yuexiang Li, Dong Wei, and Jiawei Chen contributed equally to this work.) (Corresponding author: Yefeng Zheng.)*

Yuexiang Li, Dong Wei, Jiawei Chen, Shilei Cao, Hongyu Zhou, Kai Ma, and Yefeng Zheng are with the Tencent Jarvis Lab, Shenzhen 518000, China (e-mail: vicyxli@tencent.com; donwei@tencent.com; jiaweichen@tencent.com; eliasslcao@tencent.com; hongyuzhou@tencent.com; kylekma@tencent.com; yefengzheng@tencent.com).

Yanchun Zhu, Jianrong Wu, and Tianyi Qian are with the Tencent Health, Shenzhen 518000, China (e-mail: noonezhu@tencent.com; edwinjrwu@tencent.com; tianyiqian@tencent.com).

Lan Lan, Wenbo Sun, and Haibo Xu are with the Department of Radiology, Zhongnan Hospital of Wuhan University, Wuhan 430071, China (e-mail: lanlanjulia@whu.edu.cn; sunwb3@mail2.sysu.edu.cn; xuhaibo1120@hotmail.com).

Digital Object Identifier 10.1109/JBHI.2020.3018181

[1] https://coronavirus.jhu.edu/map.html

heavy workload. Also, identifying early signs of COVID-19 in CT can be difficult for junior radiologists. Therefore, automated algorithms that are able to consistently identify COVID-19 from CT exams are of great clinical value fighting the pandemic.

In recent years, deep learning (DL) based methods have been applied to various tasks in medical image analysis and achieved outstanding performance [6], [7], including diagnosis of pediatric pneumonia using chest X-ray images [8]. Nonetheless, DL-based methods are known to be data-hungry. For example, a recent work [9] showed that the DL method was able to distinguish COVID-19 from community acquired pneumonia (CAP) and non-pneumonia on chest CT exams, using 4,356 3D chest CT exams of 3,506 patients for development and validation of the DL method, of which 1,296 (30%) were confirmed COVID-19. Under most circumstances, however, it is difficult to collect a well-balanced dataset to train DL-based models for an emerging disease like COVID-19 for most medical centers, especially in the early stage of the outbreak. On one hand, the number of positive cases (e.g., COVID-19) is extremely low. On the other hand, it is more likely that these medical centers house a considerable archive of other cases such as CAP and non-pneumonia CT exams. This results in two practical challenges to the fast development of DL-based screening methods for rapid response to disease screening and control: insufficient positive training samples and seriously imbalanced classes.

In this work, we propose a novel solution to this dilemma, for effective and efficient training of DL-based methods to aid the COVID-19 diagnosis on chest CT exams—more specifically, to distinguish CT exams of COVID-19 from others. We have collected 305 CT exams for COVID-19 and 2,370 for CAP and non-pneumonia, resulting in a 1:8 positive to negative ratio. The common strategy for handling the imbalanced training data is either to under-sample the majority class or over-sample the minority class. However, the former may discard the potentially useful information in the majority class, whereas the latter may unnecessarily increase the computational overhead. Another line of solutions is to resample [10] or reweight [11] the training samples after forwarding them through the network, especially for those in the majority class. Notwithstanding, these solutions still require loss computation for all samples in each training epoch.

Different from these methods, we propose a dual-track ranking approach to explicitly measure the 'value' of negative samples for network training in an offline manner. The proposed approach ranks the negative samples in both terms of 'difficulty' and 'diversity'. A novel self-supervised learning method (i.e., Rubik's cube Pro) is proposed to extract features from the COVID-19 and negative CT volumes. Then, two kinds of soft-labels (i.e., 'difficulty' and 'diversity') are generated for each negative sample by calculating the earth mover's distance (EMD) between its features and those of COVID-19 volumes. Using these generated soft-labels, the data 'value' of negative samples can be quantitatively evaluated. A pre-set number of negative samples are selected based on the joint ranking of 'difficulty' and 'diversity,' and fed to the neural network for training together with the COVID-19 volumes (Fig. 1).

In summary, our contributions in this work are three folds:

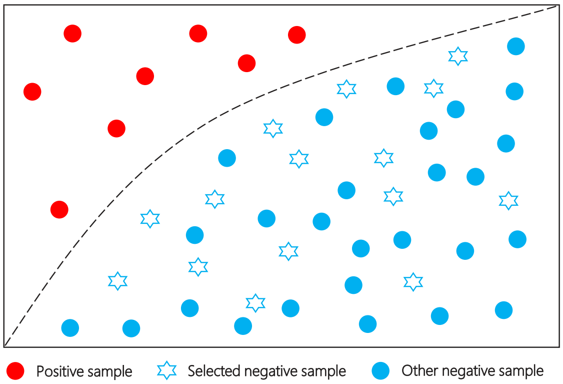

Fig. 1. Imbalanced distribution of positive (red) and negative (cyan) samples. The proposed approach is encouraged to select a subset of samples (cyan stars), which are representative of the entire negative pool (cyan stars plus cyan circles). Therefore, we can maintain the performance of classifier and shorten the training period.

- We propose a novel self-supervised learning approach, namely Rubik's cube Pro, which is an extension of the approach proposed in our previous conference paper [12]. Compared to the common Rubik's cube [12], a random masking operation is added to the pretext task to improve self-supervised feature representation, which is more robust to deal with the multi-centre CT volumes used in this study.
- A dual-track offline data selection approach is proposed to mine the informative negative samples for neural network training, which measures the data 'value' in two terms—'difficulty' and 'diversity'. Compared to the online data selection and oversampling approaches, our offline method can significantly decrease the training period of deep learning models, which is crucial for the development of computer aided diagnosis system at the early stage of the COVID-19 outbreak.
- We conduct extensive experiments to validate the efficiency and effectiveness of the proposed offline data selection approach. The experimental results show that, compared to the vanilla network training scheme using all available negative samples, our approach can achieve superior performance using only about half of the negative samples for training the COVID-19 classification network, substantially reducing model training time.

II. Related Work

In this section, we provide a systematic review of the literature related to our work. First, we review existing works on imaging-based diagnosis of pneumonia, including several concurrent works on COVID-19 diagnosis. Second, we review several topics closely related to our novel approach, including common strategies for imbalanced data, deep metric learning, and self-supervised training.

A. Imaging-Based Diagnosis of Pneumonia

A large body of literature focused on pneumonia diagnosis on chest X-ray images (e.g., [8], [13]–[16], to name a few).

Notably, both Kermany *et al.* [8] and Rajarama *et al.* [14] applied DL methods to the diagnosis between bacterial and viral pneumonia in pediatric chest X-rays, which is similar to our goal of identifying COVID-19 (a type of viral pneumonia, indeed) from pneumonia of other etiologies (CAP) and non-pneumonia. For a cost-efficient solution, Zheng *et al.* [17] proposed a DL method for pneumonia diagnosis based on clinical lung ultrasound images, using CT as the gold standard. However, both X-ray and ultrasound images are insufficient for early diagnosis of COVID-19 because they may easily miss the subtle signs such as small regions of ground-glass opacities. Prior works on machine-learning driven imaging-based diagnosis of pneumonia are rare [18], until recently several concurrent works [9], [19], [20] appeared for the diagnosis of COVID-19. However, none of these works considered the problem of seriously imbalanced COVID-19 and archived chest CT data that may be encountered in practice when developing DL methods in the early stage of the outbreak.

B. Strategies for Imbalanced Data

Class imbalance is a problem with a long history in learning-based medical image analysis, and various solutions have been proposed to deal with it. Arguably, under-sampling the majority class and/or up-sampling the minority class to roughly equilibrate their instance numbers are the most straightforward and frequently used approaches. However, under-sampling with a simple random selection strategy may discard the samples potentially informative for training the DL model. Meanwhile, simply over-sampling the minority class may not increase the amount of effective information while substantially increasing the computational overhead. In other words, neither under-sampling nor over-sampling utilizes the inherent characteristics of the available training samples. Re-weighting class losses in inverse proportion to the numbers of class instances [21] is conceptually similar to the under- and over-sampling approaches, in that it only considers the instance numbers but not the properties of the data. Both the online hard example mining (OHEM) [10] and focal loss [11] tackled the class imbalance problem by estimating the difficulty levels of the training samples, thus requiring all the training samples to be processed by the current model in each training epoch. In contrast, our approach first picks the most informative negative samples from the big archive, and only uses the selected subset for more efficient training while achieving comparable performance. The sample selection is driven by sample difficulty and diversity, both of which are estimated by a model trained via metric learning.

C. Deep Metric Learning

Metric learning has been adopted in wide computer vision applications (see [22] for a survey). Recently, this concept has been revitalized in the realm of DL and became one of the central topics in few-shot learning [23]–[25]. In the general concept of metric learning, a learnable embedding function projects the input into an embedding space, where the projected samples of the same class should be closer to each other than those from different classes. To learn the embedding function, a distance metric (most often the squared Euclidean [23] or cosine distances [25]) is employed to measure the intra- and inter-class distances, and loss functions are defined accordingly (e.g., the triplet loss [26]) to encourage small intra-class and large inter-class distances. Unlike the typical deep metric learning, we learn the embedding using only a single class (the COVID-19) through a proxy task for self-supervised training. In addition, we employ the earth mover's distance [27] rather than the usual cosine or squared Euclidean distance due to the fact that EMD is considerably more robust especially in the task of comparing two histogram-based feature descriptors.

Specifically, the earth mover's distance is a distance measure between two sets of weighted objects or distributions, which is built upon the basic distance between individual objects. It has the form of the well-studied transportation problem from linear programming. Suppose that a set of suppliers are required to transport goods to a set of demanders. The goal of EMD is to find a least-expensive flow of goods from the suppliers to the demanders. In this study, EMD is adopted to measure the similarity between two histogram-based features. The calculation process is presented in Section IV C.

D. Self-Supervised Training

To deal with the deficiency of annotated data, researchers attempted to exploit useful information from the unlabeled data with unsupervised approaches [28], [29]. More recently, the self-supervised learning, as a new paradigm of unsupervised learning, attracts increasing attentions from the community. The pipeline consists of two steps: 1) pre-train a convolutional neural network (CNN) on a proxy task with a large unannotated dataset; 2) fine-tune the pre-trained network for the specific target task with a small set of annotated data. The proxy task enforces neural networks to deeply mine useful information from the unlabeled raw data, which can boost the accuracy of the subsequent target task with limited training data. Various proxy tasks had been proposed, which include grayscale image colorization [30], jigsaw puzzle [31], image inpainting [32], spatial position prediction [33], and object motion estimation [34].

For the applications with medical data, researchers took some prior-knowledge into account when formulating the proxy task. Zhang *et al.* [29] defined a proxy task that sorted the 2D slices extracted from the conventional 3D CT and magnetic resonance imaging (MRI) volumes, to pre-train the neural networks for the fine-grained body part recognition (the target task). Spitzer *et al.* [28] proposed to pre-train neural networks on a self-supervised learning task, i.e., predicting the 3D distance between two patches sampled from the same brain, for the better segmentation of brain areas (the target task). Zhuang *et al.* [12] proposed to pre-train 3D networks by playing a Rubik's cube game, which can be seen as an extension of 2D jigsaw puzzles [35]. Zhou *et al.* [36] formulated a content restoration pretext task for 3D medical volumes, which achieved impressive improvements on multiple medical image processing tasks. In this study, we extend the Rubik's cube approach by adding a random masking operation, which improves the robustness of the self-supervised feature representation.

TABLE I
IMAGING PROTOCOLS OF THE CT EXAMS

	No. slices[a]	Slice thickness (mm)	Matrix size	Pixel spacing (mm)[b]	KVP (kV)[c]	Collimation (mm)	Pitch
COVID-19	38–629	0.625–7.5	512×512	0.604–0.977	100–130	(4–128)×(0.6–3.75)	0.75–1.5
CAP	24–1,309	0.5–7	512×512	0.482–0.951	110–140	(8–128)×(0.6–1.25)	0.734–1.438
Non-pneumonia	24–879	0.75–10	512×512	0.5–0.977	100–140	(4–128)×(0.6–2.5)	0.75–1.5

[a] The CT volumes are not of a uniform size, which makes the No. slices vary in a range.
[b] Isotropic pixel sizes.
[c] Automatic tube current was used.

III. Materials

A. Study Cohorts

This retrospective study was approved by the Institutional Review Boards of the participating hospitals, and informed consent was waived due to the retrospective nature of the study. It included 2,675 3D volumetric non-contrast thoracic CT exams from 2,595 patients acquired at over 10 medical centers between Nov. 11th, 2010 and Feb. 9th, 2020. The inclusion criteria included: (i) complete CT scans covering the whole lung, and (ii) acceptable image quality not affecting analysis. For CT exams with multiple reconstruction kernels at the same imaging session, we manually picked the one that was visually sharper. CT exams of the same subject acquired at different time points were all included. The dataset comprised 305 COVID-19 (11%), 872 CAP (33%), and 1,498 non-pneumonia (56%) exams. The data were anonymized before analysis.

All the COVID-19 data were confirmed RT-PCR positive cases, and further verified by experienced radiologists to confirm radiographical lung infections. The 305 COVID-19 CT exams were acquired on 251 patients from Jan. 18th, 2020 to Feb. 9th, 2020, using 18 scanner models made by four manufacturers (GE Medical Systems, MinFound Medical Systems, Philips, and Siemens). The CAP data comprised 872 CT exams acquired on 869 patients from Nov. 14th, 2010 to Nov. 4th, 2019, using nine scanner models made by four manufacturers (GE Medical Systems, Philips, Siemens, and Toshiba). Among them, 497 of the CAP exams were acquired at the same hospital as 80 of the COVID-19 exams. The non-pneumonia data comprised 1,498 CT exams acquired on 1,475 subjects between Nov. 11th, 2010 and Jul. 2nd, 2019, using 18 scanner models made by four manufacturers (GE Medical Systems, Philips, Siemens, and Toshiba). The non-pneumonia cohort included subjects without any lung disease or with lung nodules (with and without complications).

B. Imaging Protocol

The CT examinations were performed using various scanners made by different manufacturers, presenting considerable variation in the imaging protocol (details presented in Table I). The exams were performed with the subjects instructed to hold the breath at full inspiration.

IV. Method

A. Problem Formulation

There are two main challenges we encountered while developing the computer aided diagnosis system for COVID-19 using CT scans: 1) the quantity of COVID-19 CT volumes is deficient, compared to the available negative samples (i.e., non-pneumonia and CAP), suffering from the problem of class imbalance; and 2) due to the outbreak of COVID-19, there is limited time for the training and tuning of neural networks. Specifically, assuming we have N COVID-19 and M negative CT volumes, where $M \gg N$, the neural network trained with such an imbalanced dataset may yield biased prediction. To alleviate the problem, many approaches have been proposed such as focal loss [11] and OHEM [10]. However, those online data selection approaches are extremely time-consuming, since they take the entire sample set $(M + N)$ as the training set. Witnessed the fast outbreak of COVID-19, the tremendous time consumption caused by the online data selection approaches is unacceptable for the development of computer aided diagnosis systems.

To simultaneously address these two problems, we propose a dual-track ranking approach to explicitly measure the ‘value’ of negative samples for network training in an offline manner. As shown in Fig. 2, the proposed approach ranks the negative samples in both terms of ‘difficulty’ and ‘diversity’ by three steps. First, the proposed approach extracts features from the COVID-19 and negative CT volumes using a self-supervised learning method (i.e., Rubik’s cube Pro). Then, a soft-label is generated for each negative sample by calculating the earth mover’s distance (EMD) between the positive and negative features. Using these generated soft-labels, the data ‘value’ of negative samples can be assessed in two tracks (i.e., ‘difficulty’ and ‘diversity’). A pre-set number $(S \ll M)$ of negative samples are selected based on the joint ranking of two-tracks and fed to the neural network for training together with the positive volumes.

B. Feature Extraction via Self-Supervised Learning

An improved version of the recently proposed self-supervised approach, namely Rubik’s cube Pro, is adopted to extract discriminative features from CT volumes. Similar to [12], the 3D medical volume (original state) is seen as a $2 \times 2 \times 2$ Rubik’s cube and accordingly partitioned into eight cubes. The cubes are then processed by a series of operations (i.e., rearrangement and rotation), which formulates a disarranged state of Rubik’s cube. The aim of the Rubik’s cube pretext task is to recover the original state of medical volume from the disarranged stage. To complete the pretext task, the 3D neural network is required to deeply mine the 3D anatomical information from the raw volume. In addition to the cube rearrangement and reorientation proposed in [12], our Rubik’s cube Pro randomly masks the content of cube to increase the difficulty of the pretext task,

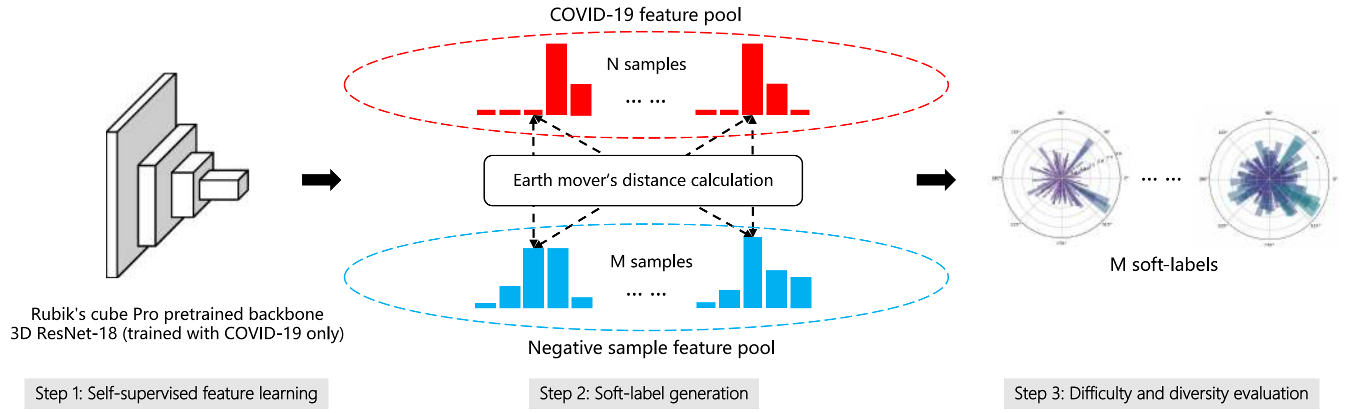
COVID-19 feature pool
N samples
... ...
Earth mover's distance calculation
M samples
... ...
Negative sample feature pool
Rubik's cube Pro pretrained backbone
3D ResNet-18 (trained with COVID-19 only)
... ...
M soft-labels
Step 1: Self-supervised feature learning
Step 2: Soft-label generation
Step 3: Difficulty and diversity evaluation

Fig. 2. The pipeline of our dual-track ranking approach. The approach involves three steps: self-supervised feature learning, soft-label generation, and difficulty and diversity evaluation. The soft-labels are presented with polar histograms, where 0° is the first element.

which encourages the neural network to learn a more robust feature representation. The proposed Rubik's cube Pro involves three operations—cube permutation, cube rotation, and cube masking, which are introduced in details in the following.

1) Cube Permutation: Taking a pocket cube, i.e., $2 \times 2 \times 2$ Rubik's cube, as an example, we first select $K = 100$ permutations from the permutation pool ($\mathcal{P}$), i.e., $\mathcal{P} = (\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_n,\ n = 8!)$, by maximizing the Hamming distance between each pair of permutations. Then, for each time of Rubik's cube recovery, the eight cubes are rearranged according to one of the K permutations. To properly reorder the cubes, the network is trained to identify the selected permutation from the K options, which can be seen as a classification task with K categories. Assuming the $1 \times K$ network prediction as z and the one-hot ground-truth as l, the permutation loss ($\mathcal{L}_P$) in this step can be defined as:

$$\mathcal{L}_P = -\sum_{j=1}^{K} l_j \log z_j. \tag{1}$$

2) Cube Rotation: Additionally, the Rubik's cube Pro pretext task adopts the random cube rotation to encourage the network to extract rotation-invariant features. To reduce the complexity of the task, we limit the directions for cube rotation, i.e., only allowing 180° horizontal and vertical rotations. To orientate the cubes, the network is required to recognize whether each of the input cubes has been rotated. It can be seen as a multi-label classification task using the $1 \times C$ (C is the number of cubes) ground truth (g) with 1 on the positions of rotated cubes and 0 vice versa. Hence, the predictions of this task are two $1 \times C$ vectors (h and v) indicating the possibilities of horizontal (hor) and vertical (ver) rotations for each cube. The rotation loss ($\mathcal{L}_R$) can be written as:

$$\mathcal{L}_R = -\sum_{i=1}^{C} (g_i^{hor} \log h_i + g_i^{ver} \log v_i). \tag{2}$$

3) Cube Masking: Inspired by the observation that learning from a harder task often leads to a more robust feature representation [37], [38], apart from the cube permutation and rotation proposed in [12], an additional operation, i.e., cube masking, is adopted in our Rubik's cube Pro to increase the difficulty of Rubik's cube recovery. Such a strategy can be seen as adding noise to the training data to avoid network overfitting. Many researches [39], [40] have proven the effectiveness of this strategy, but few of them integrate it into a self-supervised learning framework.

The process of random masking strategy can be summarized as following. First, a random possibility ($pos \in [0, 1]$) is generated for each cube to determine whether it should be masked or not. If the cube is to be masked (i.e., $pos \geq 0.5$), we generate a 3D matrix R of the same shape of a cube to randomly block the content. The value of each voxel of R is a probability ($prob$) randomly captured from a uniform distribution [0, 1]. To obtain the mask, a thresholding operation ($th_R = 0.5$) is performed to R, which leads the voxel value of (x, y, z) to be 1 for $prob(x, y, z) \geq th_R$ and 0 vice versa, where (x, y, z) is the 3D coordinate of a voxel. Assuming a cube after rearrangement and rotation as Q^o, its masking state (Q^m) can be generated by multiplying the Q^o and mask R in pixel-wise manner: $Q^m(x, y, z) = Q^o(x, y, z) \cdot R(x, y, z)$. Since the objective of our Rubik's cube Pro is the same to [12] (i.e., $\mathcal{L}_P + \mathcal{L}_R$), the neural networks recovering such partially masked Rubik's cubes can learn feature representations with robustness against disturbance of multi-centre data, compared to the original Rubik's cube.

C. Soft-Label Generation

After self-supervised learning, the pre-trained 3D neural network is then applied to extract features from both COVID-19 and negative samples (i.e., the features from the last convolutional layer of 3D ResNet-18). (Note that, to measure the distance from negative samples to the positive pool, only COVID-19 CT volumes are employed to pre-train the neural network for feature extraction.) We calculate the EMD between the features of each pair of positive and negative samples to construct an N-dimensional soft-label for each sample in the negative pool. The EMD is an actively used cross-bin distance metric due to its robustness, especially in the task of comparison between two histogram-based descriptors, such as the extracted features in our application. Denote the features of COVID-19 and negative CT samples as F_i^{COV} $(i \in \{1,\ldots,D\})$ and F_j^{NEG} $(j \in \{1,\ldots,D\})$, respectively, where D is the number of elements contained in the extracted feature ($D = 512$ of the last convolutional layer of ResNet-18). Referring to Section II C, the features of COVID-19 and negative CT samples can be seen as the sets of suppliers and demanders. The EMD between these two features can be defined as:

$$EMD(F^{COV}, F^{NEG}) = \min_{\{f_{ij}\}} \frac{\sum_{ij} f_{ij} d_{ij}}{\sum_{ij} f_{ij}} \tag{3}$$

subjected to the constraints:

$$\sum_{j}^{D} f_{ij} \leq F_i^{COV}, \ \sum_{i}^{D} f_{ij} \leq F_j^{NEG},$$

$$\sum_{i,j} f_{ij} = \min\left(\sum_i F_i^{COV}, \sum_j F_j^{NEG}\right), \quad f_{ij} \geq 0 \tag{4}$$

where f_{ij} represents the amount of flow transported from element i to j, and d_{ij} represents its ground distance. The cost for transporting the unit amount from one element to another is determined by its ground distance. The EMD is obtained by measuring the minimum total cost to transform one feature into the other. Please refer to [41] for more details.

D. Difficulty and Diversity Evaluation

Using the generated soft-labels, we can explicitly evaluate the negative samples in terms of 'difficulty' and 'diversity' to the neural network.

1) Difficulty: Since the soft-label represents the distance from the negative sample to the COVID-19 feature pool, the sample with a smaller distance is perceptually more similar to the COVID-19. Hence, the average of EMD soft-label (L) can be seen as a metric measuring the difficulty (R^{dif}) of each negative sample for the neural network to process, which can be defined as:

$$R_x^{dif} = f_{ave}(L_x) \tag{5}$$

where $f_{ave}(.)$ is the average function and L_x is the N-dimensional soft-label for sample x.

2) Diversity: The diversity of training set directly influences the robustness of neural network—a more diverse training set may lead to more robust model performance. Therefore, beyond the 'difficulty,' we further measure the 'diversity' of negative samples using the generated soft-labels. For a sample x, we first calculate the EMD between L_x and $L_m, \forall m \in [1, M]$ where $m \neq x$, which forms an $M-1$ dimensional EMD histogram. As the smaller EMD means the higher similarity of two features, the average of the obtained $M-1$ EMD histograms can represent the overall dissimilarity of the negative sample to the other negative samples—the larger the averaged EMD, the higher diversity value of the negative sample. Therefore, the 'diversity' (R^{div}) of sample x can be written as:

$$R_x^{div} = f_{ave}(EMD(L_x, L_m)), \forall m \in [1, M] \text{ and } m \neq x \tag{6}$$

where M is the number of negative samples.

3) Joint Ranking: To comprehensively assess the data 'value,' we reorder the ranking of negative samples by taking both R^{dif} and R^{div} into account. Let $f_{ind}(.)$ represent the rank indexing function (taking an integer value starting from 1), the joint ranking (R^{jnt}) can be defined as:

$$R_x^{jnt} = \alpha f_{ind}(\downarrow R_x^{dif}) + \beta f_{ind}(\uparrow R_x^{div}) \tag{7}$$

where α and β are the weights for the two tracks (empirically set to 0.5 in our experiments); and $\downarrow$ and $\uparrow$ express the operation of ascending and descending ordering. The smaller R^{dif} means higher difficulty, which is vice verse for R^{div}; therefore, the two tracks are ordering in different directions.

E. Sample Selection

The obtained joint ranking (R^{jnt}) simultaneously measures the 'difficulty' and 'diversity' of negative samples, which can be used as guideline for negative sample selection for neural network training. There are many strategies for sample selection, such as hard example mining—assigning higher weights for the difficult samples. In our experiments, to involve a variety of negative samples for the robust feature representation learning, we apply a 'long-tail' distribution sampling [42] to the $\downarrow R^{jnt}$, where a high-frequency sampling is performed to the top-ranked population, followed by a 'tails off' asymptotically sampling for the rest.

F. Implementation Details

All the experiments are conducted with the PyTorch package in Python 3.6.4 environment, using one Tesla V100 GPU.

1) Rubik's Cube Pro: Our Rubik's cube Pro is trained with a mini-batch size of 8. The initial learning rate is set to 0.001. The Adam solver [43] is used for network optimization. In our experiments, the 3D ResNet-18 [44] is adopted as the backbone. To reduce the time consumption of pretext task training, our Rubik's cube Pro is pre-trained with COVID-19 volumes only. The whole procedure of pre-training consumes about 20 minutes.

2) COVID-19 Classification Network: Straightforward data preprocessing is employed. First, we use an inhouse developed, DL-based algorithm to segment the lung masks for the 3D non-contrast CT volumes. The algorithm is robust and able to yield consistently high-quality lung masks even in the presence of severe pathological changes of the lungs. Next, we resample the CT volumes to have the unified voxel size of $1\times1\times5$ mm^3. (We have conducted primitive experiments with thin-slice resampling of 1 mm thickness, which was exponentially slower but yielded similar performance.) Then, the CT volumes are cropped according to the lung masks, adding the buffers of 9 mm, 9 mm, and 5 mm in the x, y, and z directions both pre- and post the lung masks. The lung window setting with the window level set to -600 Hounsfield unit (HU) and window width set to 1500 HU is applied to the cropped CT volumes,

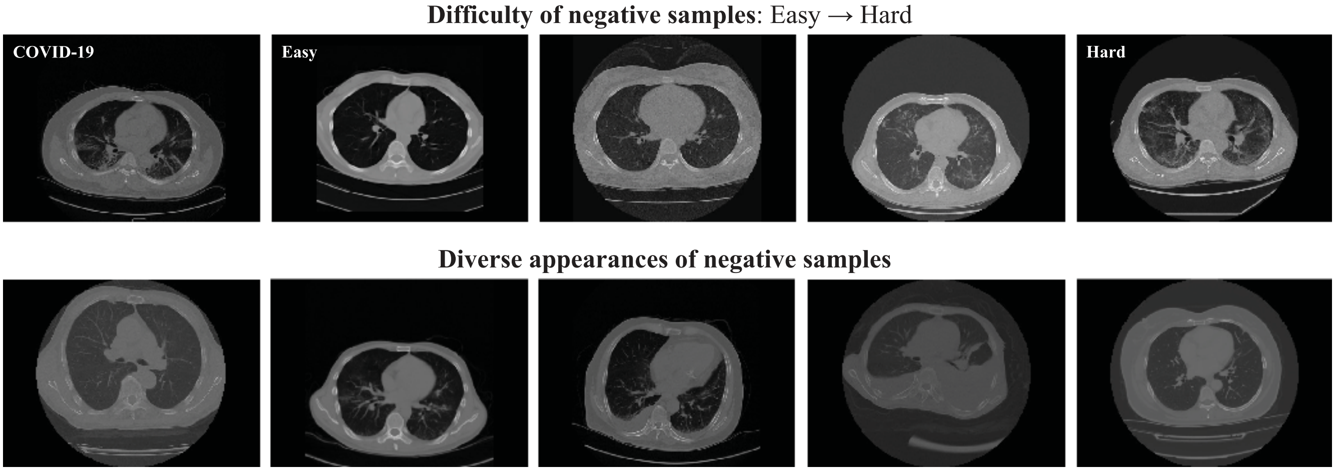
Difficulty of negative samples: Easy → Hard
COVID-19
Easy
Hard
Diverse appearances of negative samples

Fig. 3. Exemplar samples of different levels of difficulty (the first row) and diversity (the second row) selected from the dataset. The COVID-19 sample is presented in the first row for comparison. The difficulty of the samples measured by our R_{dif} gradually increases from the easy case to the hard case. The appearances of the hard samples are visually close to the COVID-19. The second row presents the negative samples with diverse appearances, which demonstrate the necessity of taking R_{div} into consideration.

whose voxel intensity values are subsequently rescaled to the range [0,1] linearly. Lastly, regions outside the lung masks are set to 0 value to eliminate interference from non-lung regions.

For input to the network, a center crop of $256 \times 192 \times 56$ voxels is taken from the preprocessed CT volume; a constant padding is performed around the volume for smaller lungs when needed. The input size is determined according to the mean size of the preprocessed CT volumes. To efficiently exploit the limited training data, two online data augmentation operations are performed during the training process, which acts as the regularization for our models: random intensity-jittering with the scale range in [0.9, 1.1] and random flipping. No data augmentation is performed for model testing.

The COVID-19 classification network adopts the same backbone and training protocol as Rubik's cube Pro. However, since the Rubik's cube Pro only uses the COVID-19 volumes for pre-training, which results in a biased network initialization, the COVID-19 classification network randomly initializes its weights, instead of using the Rubik's cube Pro pre-trained weights, at the beginning of network training. We notice that pre-training Rubik's cube Pro with all samples may boost the COVID-19 classification. However, the time consumption of pre-training increases to more than three hours, due to much more training samples (above 2,000 vs. 246). In consideration of development speed, the random initialization is preferred at the early stage of the development of computer aided diagnosis system.

V. Experiments

Due to the rapid outbreak of COVID-19, the main concerns of this study fall into two folds—the classification accuracy and training duration of the deep learning network. In this section, we conduct extensive experiments to validate the effectiveness of our offline negative sample mining approach in both terms.

TABLE II
Detailed Information of Data Split (The Number of CT Volumes)

	Training	Validation	Test
Positive	243	12	50
Negative	1,785	117	468

A. Data Split

Examples of COVID-19, CAP (the hard sample) and non-pneumonia (the easy sample) from our dataset are presented in Fig. 3, respectively. According to patient IDs, the dataset is separated to training, validation, and test sets with the ratio of 75:5:20. The number of CT volumes contained in each set is listed in Table II.

B. Evaluation Criterion

The F1 score, precision (PRE.), and recall (REC.) are adopted as the metric to evaluate the classification performance, which can be written as:

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}, \tag{8}$$

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}, \tag{9}$$

where TP, FP, and FN represent the numbers of true positive, false positive, and false negative, respectively.

C. Difficulty and Diversity Samples Visualization

To visually compare the difficulty and diversity of different negative samples, we visualize samples of different levels of difficulty and diversity selected by our approach in Fig. 3. It can be observed that the hard cases (high difficulty) on the first row are visually close to the COVID-19 sample. Oppositely, the lung texture of the easy cases (low difficulty) is clear, which

TABLE III

Ablation Study of Our Two-Track Data Selection Approach. No Online Data Selection Approach is Adopted. (PRE.–Precision, REC.–Recall (%)) *The Training Period Using Offline Data Selection Approach Includes the Time Consumption of Self-Supervised Learning (0.3 Hour)

R^{Dif}	R^{Div}	Number of Negative Samples											
		14% (246)			28% (500)			56% (1,000)			ALL (1,785)		
		PRE.	REC.	F1	PRE.	REC.	F1	PRE.	REC.	F1	PRE.	REC.	F1
-	-	**81.0**	68.0	73.9	68.8	88.0	77.2	82.0	82.0	82.0	81.5	88.0	84.6
√	-	65.7	**88.0**	75.2	72.1	88.0	79.3	81.1	**86.0**	83.5	-	-	-
-	√	79.2	76.0	77.6	**81.6**	80.0	80.8	92.9	78.0	84.8	-	-	-
√	√	74.5	82.0	**78.1**	74.2	**92.0**	**82.1**	**93.0**	80.0	**86.0**	-	-	-
w/ Pre-training		60.3	88.0	71.5	65.2	90.0	75.6	74.6	88.0	80.7	-	-	-
Training period*		2.5 hour			4.6 hours			9.5 hours			19.6 hours		

are easy for the classifier to identify. The samples with diverse appearances are presented in the second row of Fig. 3. To handle such a diverse dataset, R_{div} is a useful criterion to tune the overall diversity of the set of selected samples.

D. Ablation Study

An ablation study is conducted to evaluate the contribution made by each ranking track of our approach. The evaluation result is presented in Table III. Compared to the random selection (i.e., the first row of Table III), the neural network trained with R^{dif} or R^{dif}-based selection approache achieves consistent improvements on F1 scores with different ratio of negative samples. The approach using the joint ranking ($R^{dif} + R^{dif}$) boosts the classification accuracy with the largest margin, which are $+2.9\%$, $+2.8\%$, and $+2.8\%$ higher than the random selection with 14%, 28%, and 56% negative samples, respectively.

1) Quantity of Negative Samples: We investigate the trade-off between training period and classification accuracy by training the backbone network with different quantities of negative samples, which are 14%, 28%, and 56%. The corresponding training periods are also presented in Table III. *Note that we use a Tesla V100 GPU in the experiments; therefore, the training period will be much longer under the same setting with some lower-speed GPUs such as Tesla P40.*

As shown in Table III, fewer negative samples lead to the faster training iteration and network convergence, which shortens the period of training. However, the classification accuracy is observed to degrade in such a situation. Since our two-track data selection approach can comprehensively measure data 'value,' the neural network trained with the selected valuable negative samples achieves comparable classification accuracy to the one using the whole dataset (i.e., all 1,785 negative volumes). Furthermore, the neural network trained with 56% selected negative samples is observed to slightly outperform the one using all negative samples on precision and F1 score, which demonstrates that our approach can accurately measure the data 'value' for neural network and accordingly shorten the training period (i.e., a reduction of 52% of training time compared to the one using all samples).

2) Benefit for Shortening Development Period: We notice that the time reduction for training 3D ResNet-18 with 56% negative samples is about 10 hours compared to using all samples, which may not be significant compared to the other steps of computer aided diagnosis system development, such as backend (and possibly frontend) development. However, it is worthwhile to mention that the development of computer aided diagnosis system usually requires numerous experiments on testing different network architectures for the backbone selection, which are extremely time-consuming. In this regard, to demonstrate the generalization of the samples selected using the 3D ResNet-18, we adopt the selected samples to train a 3D ResNet-50 and evaluate its performances on the test set. The experimental results are presented in Table IV.

TABLE IV

COVID-19 Classification Accuracy of 3D ResNet-50 Trained With Different Quantities of Training Samples Selected by 3D ResNet-18. (PRE.–Precision, REC.–Recall (%); T. P.: Training Period)

	Number of Negative Samples			
	14% (246)	28% (500)	56% (1,000)	ALL
PRE.	80.9	90.2	83.0	84.0
REC.	76.0	74.0	88.0	84.8
F1	78.4	81.3	85.4	84.8
T. P.	2.9 hours	5.4 hours	11.3 hours	23.5 hours

As shown in Table IV, the training period of 3D ResNet-50 (23.5 hours) is longer than 3D ResNet-18 (19.5 hours), due to the larger quantity of network parameters. It can be observed that the ResNet-50 trained with 56% negative selected samples achieves a consistently better F1 score (85.4%) than the one using all samples (84.8%), which demonstrates the generalization and effectiveness of the selected samples. To this end, as the number of testing network architectures increases during the process of backbone selection, the benefit of shortening development period provided by our offline data selection approach gradually becomes more significant.

3) Usage of Rubik's Cube Pro Pre-Trained Weights: Common self-supervised learning approaches use the pre-trained weights as the network initialization for the subsequent target task. To this end, we train a 3D ResNet-18 initialized with the Rubik's cube Pro pre-trained weights and present its result on test set for comparison. To save the time of self-supervised learning, only the COVID-19 volumes are involved for the Rubik's cube

TABLE V
ACCURACY OF COVID-19 CLASSIFICATION NETWORK WITH DIFFERENT TRAINING STRATEGIES. (PRE.–PRECISION, REC.–RECALL (%))

	Number of Negative Samples								
	14% (246)			28% (500)			56% (1,000)		
	PRE.	REC.	F1	PRE.	REC.	F1	PRE.	REC.	F1
Random	**81.0**	68.0	73.9	68.8	88.0	77.2	82.0	82.0	82.0
Normal distribution	77.6	76.0	76.8	70.8	**92.0**	80.0	85.7	84.0	84.8
Top-R^{jnt}	41.8	66.0	51.2	60.0	78.0	67.8	84.6	**88.0**	**86.3**
Long-tail distribution	74.5	**82.0**	**78.1**	**74.2**	**92.0**	**82.1**	**93.0**	80.0	86.0
Long-tail w/ Rubik's cube Pro*	75.9	88.0	81.5	78.9	90.0	84.1	97.7	84.0	90.3

Pro pretext task, which may lead to biased predictions. The experimental results presented in Table III verify our intuition—the precision of COVID-19 classification networks using Rubik's cube Pro pre-trained weights is much lower than the random initialized ones.

E. Choice of Sampling Strategy

In Section IV E, we employed the long-tail distribution for negative sample selection from the pool after R^{jnt} ranking. Here, we further compare three sampling strategies (i.e., normal distribution, top-R^{jnt}, and long-tail distribution). The subsets of negative samples selected according to different distributions lead to varied improvements (deteriorations in some cases) in COVID-19 classification, as presented in Table V. Compared to the normal distribution sampling, the 1,000 negative samples selected by long-tail sampling contain more difficult and diverse cases, which benefit the accurate classification of COVID-19. However, when selecting smaller subsets (i.e., 246 and 500) of negative samples, the models trained with top-R^{jnt} selected samples yield extremely low classification performances. This may be because the neural network had difficulty in learning consistent information from only those samples which are most difficult and diverse. Different from that, the subsets sampled according to the long-tail distribution contain negative samples of varying 'difficulty' and 'diversity,' and outperform the others under most settings of negative sample quantity, validating the effectiveness of our choice of the sampling distribution.

1) Pre-Training Rubik's Cube Pro With All Samples: To further exploit the benefit of self-supervised learning, we pre-train the backbone network on our Rubik's cube Pro pretext task with all training samples (1785 negative and 246 positive), and then finetune with the selected ones for the COVID-19 classification. The evaluation results are listed in the last row of Table V. It can be observed that the Rubik's cube Pro pre-trained weights consistently boost the accuracy of COVID-19 classification. However, it is worthwhile to mention that although pre-training on all samples can tackle the problem of biased predictions in the positive-sample-only Rubik's cube Pro, the time consumption of pre-training increases from 0.3 hour to more than three hours, due to the substantial increase in training samples.

In our practical application, the proposed offline data selection approach is adopted for extremely rapid response to the need for fast complete development cycle of computer aided diagnosis system, including tuning network architectures, data augmentation, and learning rate, etc, even within a day. After fulling the initial need and fixing the network architecture and hyper-parameters, the model can later adopts our Rubik's cube Pro pre-training pipeline for a better COVID-19 classification accuracy.

F. Combining Online Data Selection Strategies

Although the quantity of negative samples decreases from 1,785 to 1,000 using our offline data selection approach, the positive/negative ratio is still around 1:5. To further boost the accuracy of COVID-19 classification network, we combine the widely-used online data selection, such as OHEM [10] and focal loss [11], and over-sampling approaches to deal with the imbalanced positive/negative ratio. The accuracy of the hybrid methods is presented in Table VI, where the COVID-19 classification network trained with all samples is included for comparison.

It can be observed from Table VI that the online data selection approaches produce consistent improvements in the accuracy of the COVID-19 classification network. The OHEM yields the highest improvement in F1 score, i.e., around $+3\%$ compared to no online strategy. Although the oversampling approach achieves similar F1 score to OHEM, it enlarges the training set by oversampling the positive samples, which results in an expensive time cost for network training. Furthermore, it is worthwhile to mention that the COVID-19 classification network trained with our 1,000 selected negative samples yields higher accuracy to the one using all 1,758 negative samples, which demonstrate the effectiveness of our offline data selection approach.

VI. DISCUSSIONS AND CONCLUSION

In this work, we proposed an effective (in terms of better performance) and efficient (in terms of lower time cost) training strategy for COVID-19 classification networks using a relatively small number of COVID-19 CT exams and an archive of negative samples, empowered by self-supervised learning and informative data mining. Concretely, a novel self-supervised learning method (i.e., Rubik's cube Pro) was proposed to extract features from the COVID-19 and negative samples. Then, two kinds of soft-labels (i.e., 'difficulty' and 'diversity') were generated for the negative samples by computing the earth mover's distances between the features of the negative and COVID-19 samples. Using these soft-labels, data 'values' of the negative samples were comprehensively assessed. A pre-set number of negative

TABLE VI
COMPARISON OF CLASSIFICATION PERFORMANCE YIELDED BY NETWORKS TRAINED WITH DIFFERENT ONLINE STRATEGIES AGAINST IMBALANCED DATA. (ABBREVIATIONS: PRE.: PRECISION, REC.: RECALL (%); NEG.: NEGATIVE, AND POS.: POSITIVE; T. P.: TRAINING PERIOD)

COVID-19 classification network trained w/		**No online strategy**	**Oversampling**	**OHEM [10]**	**Focal loss [11]**
ALL Neg. (1785) + 246 Pos.	PRE.	81.5	86.3	91.3	86.0
	REC.	88.0	88.0	84.0	86.0
	F1	84.6	87.1	87.5	86.0
	T. P.	19.6 hours	34.3 hours	19.6 hours	19.6 hours
56% selected Neg. (1000) + 246 Pos.	PRE.	93.0	93.3	88.2	95.2
	REC.	80.0	84.0	90.0	80.0
	F1	86.0	88.4	89.1	87.0
	T. P.	9.5 hours	15.8 hours	9.5 hours	9.5 hours

TABLE VII
ACCURACY (%) OF COVID-19 CLASSIFICATION NETWORK WITH DIFFERENT SELF-SUPERVISED WEIGHTS

	Precision	Recall	F1
Train-from-scratch	81.5	88.0	84.6
Rubik's cube [12]	86.0	86.0	86.0
Rubik's cube Pro	89.6	86.0	87.8

samples were selected offline based on the joint ranking of difficulty and diversity, and fed to the neural network for training together with the COVID-19 samples. Experimental results show that, compared to the typical training scheme using all available negative samples, our approach achieved superior classification performance using only 56% of the negative samples for training the COVID-19 classification network, reducing model training time from 19.6 hours to 9.5 hours (a 52% reduction) for a rapid response to the disease outbreak.

1) EMD-Driven Difficulty and Diversity Based Offline Sampling: We propose to select the most informative negative samples for COVID-19 classification according to the EMD-derived difficulty (R^{dif}) and diversity (R^{div}). Table III showed a comparison of the performances using either or both of R^{dif} and R^{div}, from which we noticed several interesting patterns. When using either R^{dif} or R^{div} separately, R^{dif} consistently yielded superior recall to R^{div}, whereas R^{div} consistently yielded superior precision to R^{dif}. As expected, simultaneously using both of them produced more balanced results rendering the best F1 scores. We believe this property of our proposed offline sampling strategy can be utilized for desired adjustments to precision, recall, or F1 scores of the model.

2) Combining Online Sampling for Optimal Performance: After mining of the most informative negative samples offline, we further combined our proposed offline sampling strategy with several widely used online sampling strategies for imbalanced data for optimal performance. As shown in Table VI, the online sampling strategies all achieved varying improvements, but at vastly different time cost. Notably, OHEM achieved the best F1 scores while at the same time incurred the lowest time cost, which was optimal in terms of both effectiveness and efficiency. Therefore, our final joint sampling strategy incorporated OHEM.

3) Comparison of Different Pre-Training Weights: To demonstrate the effectiveness of our Rubik's cube Pro, we compare its performance with the conventional Rubik's cube [12]. To alleviate the influence caused by different selections of training samples, we pre-train and finetune the 3D ResNet-18 with the whole training set. The evaluation results are presented in Table VII. The self-supervised learning approaches yield consistent improvements to the COVID-19 classification accuracy, compared to the train-from-scratch. As the masking operation encourages the neural network to learn more robust feature representation, the Rubik's cube Pro is observed to outperform the conventional Rubik's cube with an improvement of $+1.8\%$ for F1 score.

4) Future Work: Our main contributions in this work focus on enabling rapid development of computer aided diagnosis systems for COVID-19 screening, given limited COVID-19 samples at the early stage of the outbreak, while maintaining a high classification performance. Therefore, we adopted a relatively small network (the ResNet-18) as our backbone. As the pandemic evolves, the focus shifts from development speed to classification accuracy. Accordingly, we will experiment with more complex backbones in the future with more COVID-19 data collected. Besides, there is a margin of improvement for the diversity criterion. To further ensure the diversity of selected samples, one of the options is to sequentially select samples, forcing the next sample to be different from the eralier selected ones. However, the sequential selection may cause an expensive computation as the number of selection increases. It is worthwhile to explore an advanced criterion balancing the diversity of selected samples and the computational cost in the future. Last but not least, the proposed offline data selection approach is currently only evaluated on the binary classification task (COVID-19 vs. negative). We will make our effort to extend the method to the multi-class classification problem in the future.

REFERENCES

[1] X. Xie, Z. Zhong, W. Zhao, C. Zheng, F. Wang, and J. Liu, "Chest CT for typical 2019-nCoV pneumonia: Relationship to negative RT-PCR testing," *Radiology*, vol. 296, 2020, Art. no. 200343.

[2] Y. Fang *et al.*, "Sensitivity of chest CT for COVID-19: Comparison to RT-PCR," *Radiology*, vol. 296, 2020, Art. no. 200432.

[3] T. Ai *et al.*, "Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: A report of 1014 cases," *Radiology*, vol. 296, 2020, Art. no. 200642.

[4] C. Huang *et al.*, "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China," *Lancet*, vol. 395, no. 10223, pp. 497–506, 2020.

[5] A. Bernheim *et al.*, "Chest CT findings in coronavirus disease-19 (COVID-19): Relationship to duration of infection," *Radiology*, vol. 295, 2020, Art. no. 200463.

[6] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, 2017.

[7] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.

[8] D. S. Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[9] L. Li *et al.*, "Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT," *Radiology*, vol. 296, 2020, Art. no. 200905.

[10] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 761–769.

[11] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[12] X. Zhuang, Y. Li, Y. Hu, K. Ma, Y. Yang, and Y. Zheng, "Self-supervised feature learning for 3D medical images by playing a Rubik's cube," in *Proc. Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, 2019, pp. 420–428.

[13] P. Rajpurkar *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," 2017, *arXiv:1711.05225*.

[14] S. Rajaraman, S. Candemir, I. Kim, G. Thoma, and S. Antani, "Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs," *Appl. Sci.*, vol. 8, no. 10, 2018, Art. no. 1715.

[15] B. Antin, J. Kravitz, and E. Martayan, "Detecting pneumonia in chest X-rays with supervised learning," *Semanticscholar.org*, 2017.

[16] O. Stephen, M. Sain, U. J. Maduh, and D.-U. Jeong, "An efficient deep learning approach to pneumonia classification in healthcare," *J. Healthcare Eng.*, vol. 2019, 2019.

[17] X. Zheng, S. Kulhare, C. Mehanian, Z. Chen, and B. Wilson, "Feature detection and pneumonia diagnosis based on clinical lung ultrasound imagery using deep learning," *J. Acoustical Soc. Amer.*, vol. 144, no. 3, pp. 1668–1668, 2018.

[18] W. Shi *et al.*, "HIV-infected patients with opportunistic pulmonary infections misdiagnosed as lung cancers: The clinicoradiologic features and initial application of CT radiomics," *J. Thoracic Dis.*, vol. 11, no. 6, 2019, Art. no. 2274.

[19] X. Xu *et al.*, "Deep learning system to screen coronavirus disease 2019 pneumonia," 2020, *arXiv:2002.09334*.

[20] O. Gozes *et al.*, "Rapid AI development cycle for the coronavirus (COVID-19) pandemic: Initial results for automated detection & patient monitoring using deep learning CT image analysis," 2020, *arXiv:2003.05037*.

[21] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*.

[22] B. Kulis, "Metric learning: A survey," *Foundations Trends Mach. Learn.*, vol. 5, no. 4, pp. 287–364, 2013.

[23] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inform. Process. Syst.*, 2017, pp. 4077–4087.

[24] L. Karlinsky *et al.*, "RepMet: Representative-based metric learning for classification and few-shot object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5197–5206.

[25] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng, "PANet: Few-shot image semantic segmentation with prototype alignment," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9197–9206.

[26] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, 2009.

[27] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *Int. J. Comput. Vis.*, vol. 40, no. 2, pp. 99–121, 2000.

[28] H. Spitzer, K. Kiwitz, K. Amunts, S. Harmeling, and T. Dickscheid, "Improving cytoarchitectonic segmentation of human brain areas with self-supervised Siamese networks," in *Proc. Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, 2018, pp. 663–671.

[29] P. Zhang, F. Wang, and Y. Zheng, "Self supervised deep representation learning for fine-grained body part recognition," in *Proc. IEEE Int. Symp. Biomed. Imag.*, 2017, pp. 578–582.

[30] G. Larsson, M. Maire, and G. Shakhnarovich, "Colorization as a proxy task for visual understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 840–849.

[31] M. Noroozi, A. Vinjimoor, P. Favaro, and H. Pirsiavash, "Boosting self-supervised learning via knowledge transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9359–9367.

[32] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature fearning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2536–2544.

[33] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1422–1430.

[34] H.-Y. Lee, J.-B. Huang, M. Singh, and M.-H. Yang, "Unsupervised representation learning by sorting sequences," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 667–676.

[35] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.*, 2016.

[36] Z. Zhou *et al.*, "Models genesis: Generic autodidactic models for 3D medical image analysis," in *Proc. Int. Conf. Med. Image Comput. Comput. Assisted Intervention*, 2019, pp. 384–393.

[37] J. Deng, A. C. Berg, K. Li, and F. F. Li, "What does classifying more than 10,000 image categories tell us?" in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 71–84.

[38] C. Wei *et al.*, "Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1910–1919.

[39] R. M. Zur, Y. Jiang, L. L. Pesce, and K. Drukker, "Noise injection for training artificial neural networks: A comparison with weight decay and early stopping," *Med. Phys.*, vol. 36, pp. 4810–4818, 2009.

[40] Z. You, J. Ye, K. Li, Z. Xu, and P. Wang, "Adversarial noise layer: Regularize neural network by adding noise," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 909–913.

[41] J. Wagner and B. Ommer, "Efficient clustering earth mover‘s distance," in *Proc. Asian Conf. Comput. Vis.*, 2010, pp. 477–488.

[42] W. J. Reed, "The Pareto, Zipf and other power laws," *Econ. Lett.*, vol. 74, no. 1, pp. 15–19, 2001.

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

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.