Accurate Screening of COVID-19 Using Attention-Based Deep 3D Multiple Instance Learning

Zhongyi Han, Benzheng Wei, Yanfei Hong, Tianyang Li, Jinyu Cong, Xue Zhu, Haifeng Wei, and Wei Zhang

***Abstract*— Automated Screening of COVID-19 from chest CT is of emergency and importance during the outbreak of SARS-CoV-2 worldwide in 2020. However, accurate screening of COVID-19 is still a massive challenge due to the spatial complexity of 3D volumes, the labeling difficulty of infection areas, and the slight discrepancy between COVID-19 and other viral pneumonia in chest CT. While a few pioneering works have made significant progress, they are either demanding manual annotations of infection areas or lack of interpretability. In this paper, we report our attempt towards achieving highly accurate and interpretable screening of COVID-19 from chest CT with weak labels. We propose an attention-based deep 3D multiple instance learning (AD3D-MIL) where a patient-level label is assigned to a 3D chest CT that is viewed as a bag of instances. AD3D-MIL can semantically generate deep 3D instances following the possible infection area. AD3D-MIL further applies an attention-based pooling approach to 3D instances to provide insight into each instance's contribution to the bag label. AD3D-MIL finally learns Bernoulli distributions of the bag-level labels for more accessible learning. We collected 460 chest CT examples: 230 CT examples from 79 patients with COVID-19, 100 CT examples from 100 patients with common pneumonia, and 130 CT examples from 130 people without pneumonia. A series of empirical studies show that our algorithm achieves an overall accuracy of 97.9%, AUC of 99.0%, and Cohen kappa score of 95.7%. These advantages endow our algorithm as an efficient assisted tool in the screening of COVID-19.**

Manuscript received May 7, 2020; accepted May 15, 2020. Date of publication May 21, 2020; date of current version July 30, 2020. This work was supported in part by the Natural Science Foundation of China under Grant 61872225, in part by the Natural Science Foundation of Shandong Province under Grant ZR2019ZD04 and Grant ZR2015FM010, in part by the Introduction and Cultivation Program for Young Creative Talents in Colleges and Universities of Shandong Province under Grant 173, and in part by the Shandong Province Major Science and Technology Innovation Project under Grant 2020SFXGFY04-1. *(Corresponding authors: Benzheng Wei; Wei Zhang.)*

Zhongyi Han, Benzheng Wei, Yanfei Hong, Tianyang Li, and Jinyu Cong are with the Center for Medical Artificial Intelligence, Shandong University of Traditional Chinese Medicine, Qingdao 266112, China (e-mail: hanzhongyicn@gmail.com; wbz99@sina.com; hongyanf10@gmail.com; lty@foxmail.com; congjinyu1991@gmai.com).

Xue Zhu, Haifeng Wei, and Wei Zhang are with the Affiliated Hospital of Shandong University of Traditional Chinese Medicine, Jinan 250014, China (e-mail: zhu-99@163.com; weihaifeng1979@163.com; huxizhijia@126.com).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2020.2996256

***Index Terms*— COVID-19, SARS-CoV-2, screening, computer-aided diagnosis, multiple instance learning, attention, 3D, deep learning, machine learning.**

I. Introduction

With the outbreak and widespread of SARS-Cov-2 worldwide, artificial intelligence (AI) assisted screening of COVID-19 from chest CT is significantly urgent and necessary. SARS-Cov-2 is a novel virus with the human-to-human transmission, causing an ongoing pandemic of the respiratory illness known as coronavirus disease 2019 (COVID-19). To date, SARS-Cov-2 has attacked 216 countries, areas, or territories that involve 6,272,098 confirmed COVID-19 cases and 379,044 confirmed deaths according to WHO. Toward fast stopping the widespread of COVID-19, large-scale screening is imperative to cut off the source of infection. Clinical practice demonstrates that chest CT is an effective inspection strategy because it can characterize the standard features between the majority of COVID-19 cases, which show ground-glass opacities in the early stage and pulmonary consolidation in the late stage [1], [2]. While nucleic acid detection of reverse transcription-polymerase chain reaction is a gold standard to screen COVID-19, the availability, stability, and reproducibility of the nucleic acid detection kits are questionable [3]. For example, some patients need to be checked repeatedly because the false-negative rate is high [3]. Chest CT seems particularly essential and even is called to replace the detection kits as one of the early diagnostic criteria in a period of time [4]. However, clinical screening of COVID-19 from chest CT is under problem with enormous pressure, and its screening sensitivity is unsatisfactory according to the screening performance test of radiologists [3]. Automated tools can correspondingly assist the clinical practice in speeding up screening and improving the sensitivity. Therefore, *automated screening of COVID-19 from chest CT*, the main topic of our analysis, is urgently needed to deal with this problem.

However, accurate screening of COVID-19 still faces enormous challenges from the *spatial complexity* of 3D volumes, the *labeling difficulty* of infection areas, and the *slight discrepancy* between COVID-19 and common viral pneumonia in

chest CT. Firstly, despite the advantages of CT compared with traditional 2D medical radiography, the volumes of CT generally include hundreds of slices that bring in more difficulties for computational analysis. Secondly, one possible automated approach for screening COVID-19 is to build a classifier on segmented infection areas, unfortunately, which are hardly labeled manually due to the necessity of expensive cost and the indistinguishable characteristic of ground-glass opacities. Finally, COVID-19 and other viral pneumonia share similar features. Even radiologists cannot distinguish them from chest CT accurately without other inspection methods [3].

While a few previous studies have made significant advancement for automated screening of COVID-19 from chest CT, they are either demanding manual annotations of infection areas or lack of interpretability. According to the input types of screening classifiers, we divide the pioneering methods into three classes. The first class is the patch-based methods that leverage a segmentation model to detect infection areas and then train a classifier based on them [3]–[6]. Although the two-stage manner is similar to the observation processes of radiologists on Chest CT, supervised segmentation needs a large scale of annotations of infection areas. Even if unsupervised segmentation algorithms do not need annotations, but they are prone to errors. The second class is the slice-based methods that use a 2D model to perform slice-wise decisions [7]–[10]. However, such an approach needs to manually select infection slices among hundreds of chest CT slices for training. The third class is the 3D CT-based method that takes 3D CT scans as input, and use 3D convolutional neural networks (CNN) to make decisions directly [11]. This direct approach can avoid mistakes caused by intermediate processes; however, it still a black-box model lacking interpretability of results. In summary, the direct yet interpretable algorithms would be more helpful and compelling; however, they are under-explored so far.

In this paper, we propose an attention-based deep 3D multiple instance learning (AD3D-MIL) approach towards achieving accurate and interpretable screening of COVID-19 from chest CT. Generally speaking, AD3D-MIL views each 3D chest CT as a bag of instances that can be interpreted into small 3D cubes. The main goal of AD3D-MIL is learning to predict an individual category label assigned to a chest CT, *e.g.,* COVID-19, common pneumonia, or no pneumonia. Another essential objective is to obtain crucial instances that can reveal the location of infection areas. Unlike previous MIL works that assume the existence of already-separated instances, AD3D-MIL could semantically generate deep 3D instances with permutation-invariance. For improving the interpretability of results, AD3D-MIL involves an attention-based MIL pooling strategy applied on deep 3D instances to give insight into every instance's contribution to the bag label. AD3D-MIL finally learns Bernoulli distributions of the bag-level labels for more accessible learning. Since existing MIL works are mainly focusing on binary classification, AD3D-MIL can also extend MIL to multi-class classification by modeling the joint Bernoulli distribution of multi-class bag labels. We seamlessly transform AD3D-MIL into a 3D neural network that performs efficient end-to-end optimization by backpropagation, successfully achieving accurate screening of COVID-19. A series of empirical studies on a newly-collected dataset show that AD3D-MIL, with interpretability of results, remarkably exceeds the state of the art works.

The main contributions of this study include: The main contributions of this study include:

- We achieve an accurate and interpretable screening of COVID-19 that contribute to the large-scale screening in clinical for the fast stopping of COVID-19 worldwide.
- In the screening problem of COVID-19, we propose a weakly-supervised learning framework that unifies attention mechanism and deep multiple instance learning in a mutually beneficial way.
- We propose, for the first time, an automated deep 3D instances generator with robust scalability and flexibility. This approach can extend the MIL into practical tasks.

We arrange the remainder of this article as follows. In Section II, we review the related works in terms of artificial intelligence assisted analysis of COVID-19 and involved methodology. We give some preliminaries in Section III and describe in detail the proposed AD3D-MIL in Section IV. In Section V, we provide detailed descriptions of collected datasets, experiment settings, and evaluation results. Finally, we conclude and discuss this study in Section VI.

II. Related Work

This section presents related works in terms of automated screening of COVID-19 and involved methods of our work.

A. Automated Screening of COVID-19

Since medical imaging plays a fundamental function in the global fight against COVID-19, lots of works have been devoted to AI-empowered technologies to improve the work efficiency of medical image analysis. These emerging works of COVID-19 includes automated screening [3]–[11], lesion segmentation [8], [12], infection quantification [13], and patient severity assessment [14]. Among them, automated screening attracts the most attention, for instance, which takes up much space in the first comprehensive review paper about AI for COVID-19 [15]. Generally speaking, pioneering screening works include chest X-ray based and chest CT based works. Chest X-ray based works leverage 2D CNNs to make decisions directly [16]–[20]. While chest X-ray has the characteristics of low radiation and low cost, chest CT is the most commonly used inspection strategy for the COVID-19 diagnosis because it can characterize the most common findings [2]. Accordingly, a large part of screening works is built on chest CT.

Due to the spatial complexity of chest CT, existing screening works attempt to handle that by adopting three different strategies. The first type is the patch-based methods. As far as we know, Xu *et al.* is the first work attempting to study the automated screening of COVID-19 from chest CT [4]. Based on 618 CT scans, they first leveraged VB-Net to extract regions of interest (ROI) and then used a CNN to screen COVID-19 from Influenza-A viral pneumonia and irrelevant to infection groups. Wang *et al.* first used a threshold approach to extract ROI images and then trained a modified inception network to screen

COVID-19 from typical viral pneumonia [3]. They collected chest CT scans from 79 cases of COVID-19 and 180 cases of typical viral pneumonia with an accuracy of 79.3%. Based on a large-scale dataset, Shi *et al.* first trained a VB-Net for the segmentation of ROIs and then extracted manually-designed features to train a random forest on classifying COVID-19 and common pneumonia [5]. Jin *et al.* attempted to combine 3D U-Net and ResNet-50 to build a screening system in four weeks with satisfying performance [6]. As we mentioned before, these two-stage approaches either need lesion annotations or are prone to the errors from intermediate steps.

The second type is the slice-based methods. Gozes *et al.* used a 2D CNN to perform slice-level classification on 270 slices comprised of 120 COVID-19 and 150 normal slices [8]. Based on a multi-center dataset comprised of 88 COVID-19, 101 bacteria pneumonia, and 86 healthy CT scans, Song *et al.* applied a modified residual network (ResNet-50) for slice-level classification [7]. Moreover, Jin *et al.* [10] and Gozes *et al.* [9] constructed a same pipeline: a deep ResNet-50/152 to perform slice-level classification and a gradient class activation mapping for show the heatmaps. Note that while the heatmaps can also explain results, they are post-hoc analyses. The slice-based methods need the manual selection of slices to train the classifier, and they neglect the spatial correlation in CT scans, which is key for the screening of COVID-19.

The last type is the 3D CT-based method, and there is only one existing work to date. Based on 540 CT scans comprised of 313 COVID-19 and 229 others, Zheng *et al.* attempted to leverage a 3D CNN to make decisions directly with satisfying performance [11]. Since a 3D model is more complicated than a 2D model, this type of resolution lacks the interpretability of results. On the other hand, this direct manner can achieve optimal minima by leveraging end-to-end optimization, which often obtains better performance than multi-stage methods. Therefore, to realize the direct screening and interpretability of results simultaneously, we propose a novel algorithm of multiple instances learning that integrates the expression ability of key instances and the end-to-end optimization.

B. Involved Methods of Our Work

1) Multiple Instance Learning (MIL): MIL is a type of inexact supervised learning that is a branch of weakly supervised learning [21]. Concretely, MIL receives coarse-grained labels where the training data is imperfect. One seminal work in this field was conducted by Dietterich *et al.* [22]. For analyzing the multiple instance setting, this work attempted three types of approaches for learning axis-parallel rectangles. It showed that the algorithm that ignores the multiple instance setting performs very unsuccessfully. After that, many powerful algorithms appeared and performed at two levels: instance-level or bag-level [23]. Since developing the instance-level classification algorithm demands ground truth of instance labels, most studies focus on the bag-level MIL setting. Almost all bag-level classification algorithms are extended from supervised learning algorithms, including MI-SVM [24], MIL-Boost [25], EM-DD [26], and MILD [27]. These algorithms consider learning an optimal classification boundary for the MIL problem.

MIL has been successfully applied to various domains over the last 20 years, such as computer-aided diagnosis and detection [28]–[32], image classification/retrieval/annotation [33]–[36], text categorization [37], spam detection [38], object detection [25], unsupervised saliency object discovery [39], object tracking [40], etc. When MIL applies to medical image analysis, the occurrence and structures of instances (organs) are beneficial for MIL classifier [23]. For example, Melendez *et al.* shown an obvious performance gain by training an MI-SVM classifier on distinguishing chest X-ray images into healthy or containing tuberculosis [31]. The apparent improvement is also obtained on the task of diagnosis of chronic obstructive pulmonary disease (COPD) from breast CT [32]. Ilse *et al.* proposed a new attention-based deep multiple instance learning framework used for the intelligent detection of cancerous regions in histopathological slides, in which ROIs can be indicated [41]. To achieve the computer-aided diagnosis of endoscopic diseases using weak labels, Wang *et al.* formulated this task as a MIL problem and built a weakly labeled endoscopic image dataset [42]. For more work, please refer to the survey given by Cheplygina *et al.* [43]. Compared with existing MIL algorithms, our AD3D-MIL mainly has four advances: 1) extent MIL to 3D tasks, 2) can generate instances automatically, 3) achieve the classification of multi-class bags, and 4) introduce attention mechanism to discover key instances that indicate the infection area of COVID-19 on chest CT.

2) Attention Mechanism With MIL: Embedding attention mechanisms in deep learning is an attempt to mimic human brain actions concentrating on a few important things. It has given birth to the rise of many breakthroughs in the field of natural language processing (NLP), such as Transformer architecture [44] and Google BERT [45]. The attention mechanism based deep learning framework is generally adopted in the fields of image captioning [46] and text analysis [47]. There are only four works integrates attention mechanism into MIL problem and are in a minimal form. Pappas *et al.* attempted to use an additional linear regression module to compute the attention weights on instances. A one-layer network then replaces the linear regression model with a single output [48]. Qi *et al.* attempted to use attention-based MIL operator for the classification and segmentation of point sets [49]. However, this attempt performed worst than the max operator of instance pooling. To improve this attempt, Ilse *et al.* proposed to adopt two fully-connected layers as a neural network to learn an attention-based MIL operator and demonstrated that this idea exceeds the max operator and the mean operator [41]. Inspired by this idea, we propose to use the attention mechanism to apply on 3D data with automated instance generation and end-to-end optimization through backpropagation.

III. Preliminaries

In this section, we present the necessary notations and objectives for the task of screening of COVID-19 from chest CT and then present the underlying assumptions and popular approaches for the problem of multiple instance learning.

A. Problem Setting

We first consider the familiar supervised learning setting in which the learner receives a sample of m labeled training examples $\{(X_i, Y_i)\}_{i=1}^m$ drawn from a joint distribution Q defined on $\mathcal{X}\times\mathcal{Y}$, where $\mathcal{X}$ is the instance set and $\mathcal{Y}$ is the label set. $\mathcal{Y}$ is $\{0, 1\}$ in binary classification and $\{1, \ldots, K\}$ in multi-class classification. Denote by $\hat{Q}$ the empirical distribution. In this task, $\mathcal{X}$ is the set of chest CT scans, and $\mathcal{Y}$ is the set of patient-level labels. X_i is any chest CT scan of one patient, and Y_i is the label of this patient.

We denote by $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a loss function defined over pairs of labels. For binary classification, we denote by $f : \mathcal{X} \to \{0, 1\}$ a scoring function, which induces a labeling function $h_f : \mathcal{X} \to \mathcal{Y}$ where $h_f : X \to \arg\max_{Y\in\mathcal{Y}} f(X, Y)$. For any distribution Q on $\mathcal{X}\times\mathcal{Y}$ and any labeling function h_f, we denote $\epsilon_Q(h_f) = \mathbb{E}_{(X,Y)\sim Q}\ell(h_f(X), Y)$ the expected risk. Our objective is to select a hypothesis f out of a hypothesis set $\mathcal{F}$ with a small expected risk $\epsilon_Q(h_f)$ on the target distribution.

B. Multiple Instance Learning

1) MIL Formulation: The MIL algorithm acquires a sample of m training examples $\{(X_i, Y_i)\}_{i=1}^m$ drawn from a joint distribution Q defined on $\mathcal{X} \times \mathcal{Y}$. Note that X_i is a bag of instances and $X_i = \{x_1, x_2, \ldots, x_N\}$ where N denotes the quantity of instances in a bag. Furthermore, we assume that each instance x_n has a individual label $y_n \in \{0, 1\}$, for $n = 1, \cdots, N$. However, in this task, these instance labels y are not easily available due to the expensive annotation cost in clinical. X_i is any chest CT scan of one patient, and Y_i is the label of this patient. Note that any instance x_i is a small volume in a CT scan, and it may involve the infection area of COVID-19.

2) MIL Assumption: Traditional MIL studies agree that the assumption of MIL is as follows.

$$Y = \begin{cases} 0, & \text{iff } \sum_n y_n = 0, \\ 1, & \text{otherwise.} \end{cases} \tag{1}$$

In our work, this assumption indicates that the chest CT is from a COVID-19 patient if it involves at least one lesion. Based on this assumption, the empirical loss is formulated by

$$\epsilon_{\hat{Q}}(h_f) = \frac{1}{m}\sum_{i=1}^{m}\ell(h_f(X_i), Y_i), \tag{2}$$

where $h_f(\cdot)$ represents a labeling function induced by an MIL scoring function f, and $\ell(\cdot, \cdot)$ can be any loss functions, such as 0-1, hinge loss, etc.

3) MIL Decomposition: In practice, the process of MIL includes several steps, in which each step corresponds to a specific transformation function. Given an input instance x_n, the whole scoring function f of the MIL problem can be revised into

$$f(X) = g(\sigma_{x_n\in X}[\psi(x_n)]), \tag{3}$$

where ψ and g are continuous functions. σ is a symmetric function, *e.g.,* max, mean. Accordingly, the MIL problem can be decomposed into three steps: 1) a transformation function ψ to obtain the features or pseudo-labels of instances; 2) an asymmetric function σ to generate the feature or predication label of the bag by combining the features or pseudo-labels of instances; 3) if the feature of the bag is generated, a transformation function g to pursue the final label of the bag. Otherwise, this step is needless.

4) MIL With Raw Instances: Traditional MIL methods do assume that the instances are pre-defined and segmented in advance. For example, each instance has been defined by the researcher, and the features of each instance have been extracted, *i.e.,* the transformation ψ is needless. However, the instances of lots of real-world tasks are raw without extracted features [28]–[32]. Owing to the strong expression ability, neural networks are used for the representation extraction of instances. Given an raw instance x_n, a neural network ψ with parameters θ_ψ transforms it into a hidden feature h_n: $h_n = \psi(x_n)$, in which $h_n \in \mathcal{R}$. Note that $\mathcal{R} = [0, 1]$ for instance-level approach, while $\mathcal{R} = \mathbb{R}^D$ for embedding-level approach. The goal of the instance-level approach is to predict the label of instances rather than generating features for them.

On the contrary, the objective of the embedding-level approach is to generate the features of raw instances. As mentioned above, the function g is needless for the instance-level approach. For the embedding-level approach, the function g can also be a neural network to make final decisions based on the representation z of the bag. The only restriction is that the symmetric function σ must be differentiable. To achieve that, MIL pooling operators are leveraged to integrate the learned representation of instances. There are two standard differentiable MIL pooling operators: the maximum operator:

$$\forall_{d=1,\ldots,D} : z = \max_{n=1,\ldots,N}\{\boldsymbol{h}_{nd}\}, \tag{4}$$

and the mean operator:

$$z = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{h}_n. \tag{5}$$

Both MIL pooling operators are viewed as neural layers and widely used in the MIL with neural networks. Note that MIL pooling is different from the max or average pooling layers of CNNs that perform on the feature maps.

5) Disadvantages: While MIL with neural networks has made substantial impacts in advancing algorithm designs, there are two crucial directions for improvement:

1) While the approaches of MIL with neural networks can extract deep features from a bag of raw instances, they also need the instances that are separated already. However, manual separation of instances is inefficient and suboptimal in many real-world tasks like video or image analysis, even for the 3D medical image analysis. The other problem caused by this approach is that when new tasks appear, the researcher still needs to separate instances. In the specific new tasks, such as the 3D CT based screening of COVID-19, the instances are hard to be defined and designed due to the difficulties of infection area labeling. As a result, the previous MIL algorithms cannot be used directly.

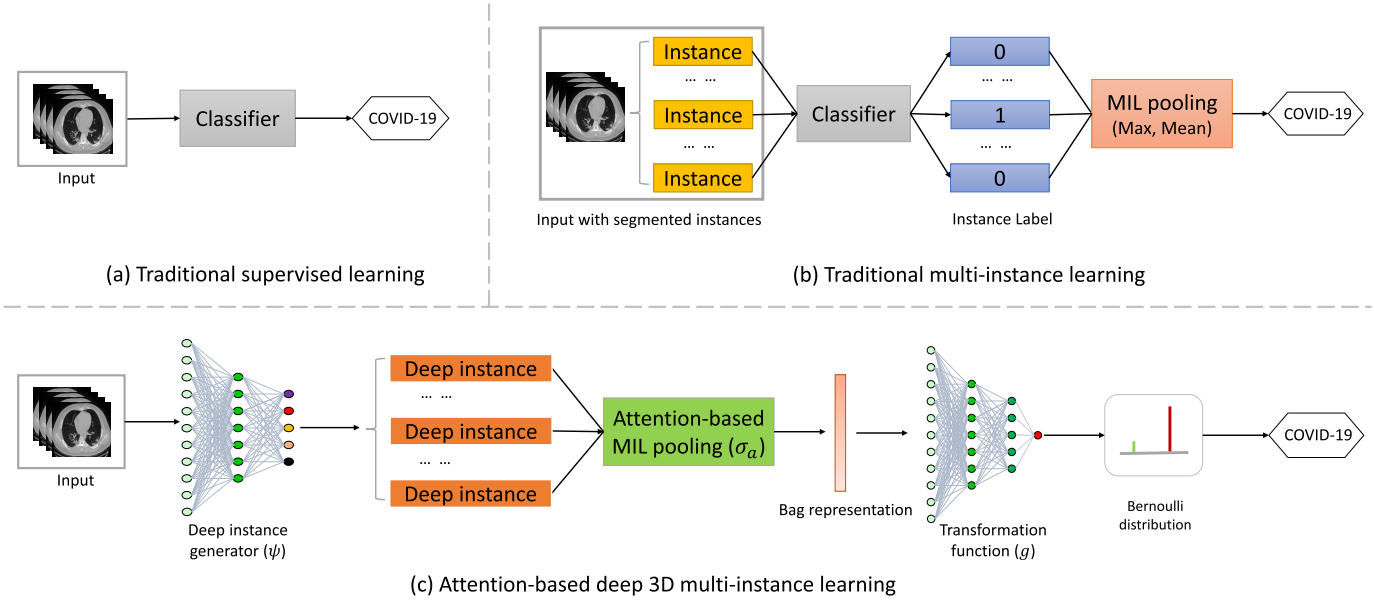


Fig. 1. An illustration of the proposed attention-based deep 3D multi-instance learning compared with traditional learning paradigms.

2) Both maximum and mean pooling operators are predefined and non-trainable. Maximum pooling is unsuitable for the embedding-level approaches, while mean pooling cannot find the key instances.

These directions are significant challenges for practical algorithm designs. In this paper, we aim to overcome these challenges by designing an attention-based deep 3D multi-instance learning.

IV. The Proposed Approach

According to the above directions, we propose the attention-based deep 3D multi-instance learning (AD3D-MIL). As illustrated in Figure 1, compared to traditional multi-instance learning, AD3D-MIL first transforms a raw unseparated bag into multiple 3D instances with semantic representation (see Section IV-A). It then combines the deep 3D instances into the bag representation using an attention-based MIL pooling (see Section IV-B). It finally transforms the bag representation into the final prediction by using a neural network to learn the Bernoulli distribution of the bag (see Section IV-C). We integrate these three steps into a 3D deep neural network for end-to-end optimization. We then relax the assumption of MIL into multi-class classification problem (see Section IV-D). Finally, we summarize the advantages (see Section IV-E).

A. Deep Instance Generation

As mentioned before, existing popular approaches of MIL with neural networks treat separated instances as inputs, then use a deep neural network to transform them into embedding space. However, such a manner neglects the spatial and global information between instances among 3D CT scans. Here we propose a deep instance generator ψ that treats one 3D CT scan as a whole and generate deep instances automatically. Generally speaking, the deep instance generator can be a fully 3D convolutional neural network (CNN). In practical, given a 3D chest CT scan X_i with the shape of $H \times W \times S$, the final layer of 3D fully CNN outputs a series of 3D feature maps with the shape of $H^* \times W^* \times S^* \times D$, where H^*, W^*, S^*, and D represent the high, width, spatial, and feature dimension of 3D feature maps, respectively.

Inspired by [50], we view each point of the $H^* \times W^* \times S^*$ cube as an instance with dimension of $D \times 1$. That is, inside the final layer of the deep instance generator, there are total $N = H^* \times W^* \times S^*$ instances generated with deep representation. Following the former notations, we can generate a bag of deep 3D instances: $\mathcal{H}_i = \{h_1, h_2, \ldots, h_N\}$ where N denotes the quantity of instances in a bag, $\mathcal{H}_i \in \mathbb{R}^{N \times D}$. Note that the raw location of corresponding instances on the 3D chest CT can be easily derived according to the location of deep instances on the cube. Formally, this step can be formulated into:

$$\mathcal{H}_i = \psi(X_i), \tag{6}$$

where X_i is a bag of raw input and $\mathcal{H}_i \in \mathbb{R}^{N \times D}$.

In conclusion, the transformation ψ in our work not only transforms of instances into embedding space but generates instances that are not defined before. Viewing each point in the feature maps as an instance is a straightforward routine to create deep 3D instances that consider the spatial relations between instances. The main difference with the existing method [50] is that our generator can apply on 3D data.

B. Attention-Based MIL Pooling

Since maximum and mean MIL pooling operators have clear disadvantages, a flexible and adaptive MIL pooling approach would be desirable for achieving hopeful performance. After obtaining a bag of deep 3D instances $\mathcal{H}$, we embed the attention-based MIL pooling approach into the AD3D-MIL framework for achieving interpretable screening of COVID-19. The attention-based MIL pooling is an interpretable symmetric function proposed by [41].

Formally, we denote by $\mathcal{H} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_N\}$ a bag of N deep instances. the attention-based MIL pooling is defined by:

$$\boldsymbol{z} = \sum_{n=1}^{N} a_n \boldsymbol{h}_n, \tag{7}$$

where,

$$a_n = \frac{\exp\{\boldsymbol{w}^\top \tanh(\boldsymbol{V}\boldsymbol{h}_n^\top)\}}{\sum_{j=1}^{N} \exp\{\boldsymbol{w}^\top \tanh(\boldsymbol{V}\boldsymbol{h}_j^\top)\}}, \tag{8}$$

where $\boldsymbol{w} \in \mathbb{R}^{N\times 1}$ and $\boldsymbol{V} \in \mathbb{R}^{N\times D}$ are trainable parameters. The hyperbolic tangent $\tanh(\cdot)$ element-wise non-linearity is used for proper gradient flow. The main difference with the existing attention mechanism is that we apply the attention mechanism on deep 3D instances. Intuitively, if a deep 3D instance is assigned to the biggest attention weight, it is the key instance. That is, the attention weights can give insight into every instance's contribution to the bag label. Therefore, the attention-based MIL pooling gives strong interpretability for the predictions. Also, the generated bag representation $\boldsymbol{z}$ is more semantic than traditional MIL pooling operators. In summary, let σ_a with parameters θ_{σ_a} represent the the attention-based MIL pooling, this step can be formulated into:

$$\boldsymbol{z}_i = \sigma_a(\mathcal{H}_i). \tag{9}$$

Based on 3D convolutional neural networks, the attention-based MIL pooling module can receive deep 3D instances and generate the semantic representation for 3D data. These endow AD3D-MIL the ability to process 3D CT data. Because 3D data contains more and more instances than 2D data, the task of the 3D MIL task is more complex and challenging than 2D data. Therefore, the setting of instance number is essential.

C. Transform Into Final Bag Labels

Given a representation $\boldsymbol{z}_i$ of a bag X_i, we use two fully connected layers as the transformation function g. This function can transform the bag representation $\boldsymbol{z}_i$ into the bag label Y_i. Specifically, this step can be formulated into:

$$Y_i = \begin{cases} 1, & \text{iff } g(\boldsymbol{z}_i) > \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

We define the distribution of the bag label to Bernoulli distribution with the parameters θ_g, *i.e.*, $g(\boldsymbol{z}_i) \in [0, 1]$, which represents the probability p_i of $Y_i = 1$ given the bag representation $\boldsymbol{z}_i$. In this paper, the Bernoulli distribution is a discrete distribution having two possible outcomes labeled by $Y = 1$ and $Y = 0$ in which $Y = 1$ (COVID-19) occurs with probability p and $Y = 0$ (Non-COVID-19) occurs with probability $1 - p$, where $0 < p < 1$. We use the two fully connected layers to learn the Bernoulli distribution of the bag label, where neural networks fully parameterize the bag label probability. The final layer outputs a scalar that represents the probability of being COVID-19. If the probability $p > \tau$ (a threshold), the bag label is COVID-19, else is Non-COVID-19. Without loss of generality, the final bag label $\hat{Y}_i$ is determined by the threshold τ of 0.5. Note that the transformation function g projects the bag representation into Bernoulli distribution rather than a binary vector generated by the traditional softmax layer. Compared with the softmax layer, such a manner is more suitable for the MIL hypothesis. It makes the learning (optimization) problem easier through learning a MIL algorithm by minimizing the log-likelihood function as follows.

Algorithm 1 AD3D-MIL Algorithm

input : parameters θ_ψ, θ_{σ_a}, θ_g, learning rate η, max epoch T, threshold τ
output: θ_ψ, θ_{σ_a}, θ_g
initialize parameters θ_ψ, θ_{σ_a}, θ_g
for t = *1, 2, ..., T* **do**
  /* step 1: Deep 3D instance generation */
  preprocess 3D CT scans $[X]_{i=1}^{m}$
  obtain feature maps:$O = \psi(X)$
  reshape feature maps O into $\mathcal{H}$
  /* stage 2: attention-based MIL pooling */
  obtain attention weight $\boldsymbol{a}$ by Equation (8)
  combine instance representation $\boldsymbol{z} = \sum_{n=1}^{N} a_n \boldsymbol{h}_n$
  /* stage 3: transform into Bernoulli distribution */
  obtain Bernoulli distribution $p = g(\boldsymbol{z})$ of bag
  produce bag label $\hat{Y}$ with threshold τ
  update $\theta_\psi = \theta_\psi - \eta\nabla\ell(\hat{Y}, Y)$
  update $\theta_{\sigma_a} = \theta_{\sigma_a} - \eta\nabla\ell(\hat{Y}, Y)$
  update $\theta_g = \theta_g - \eta\nabla\ell(\hat{Y}, Y)$
end

D. Optimization and Extension

We finally integrate the deep instance generator ψ, attention-based MIL pooling σ_a, and transformation function g into an end-to-end optimization by backpropagation. The workflow of optimization is shown in Algorithm 1. For the traditional MIL problem with binary classification, we minimize a log-likelihood loss function, which is in the following form:

$$\arg\min_{\theta_\psi, \theta_{\sigma_a}, \theta_g} -\sum_{i=1}^{m} Y_i \log(g(\sigma_a[\psi(X_i)])) + (1 - Y_i)\log(g(\sigma_a[\psi(X_i)])). \tag{11}$$

In practice, multi-class classification is demanding. For example, practical screening of COVID-19 not only needs the model to distinguish chest CT into COVID-19 and Non-COVID-19, but also demands the model to distinguish chest CT into COVID-19, common pneumonia, and no pneumonia due to the difficulty of distinguishing COVID from viral pneumonia.

Typical MIL approaches leverage one-vs.-rest (OvR) or one-vs.-all (OvA) strategies, but which need to train multiple models. In this work, we relax the assumption of MIL problem, that is, only if given a bag representation $\boldsymbol{z}_i$, we can construct a multi-class transformation function g_{mc} that projects $\boldsymbol{z}_i$ into a joint Bernoulli distribution $p_i = p_i(Y_i = 1) \cdot p_i(Y_i = 2) \cdot \cdots \cdot p_i(Y_i = K)$ where K is the class number. The class with max probability is the final label of the bag. For the MIL

problem with multi-class classification, we minimize the multi-class cross-entropy loss function without the softmax function, which is in the following form:

$$\arg\min_{\theta_\psi,\theta_{\sigma_a},\theta_g} -\sum_{i=1}^{m} p(Y_i)\log(g_{mc}(\sigma_a[\psi(X_i)]). \tag{12}$$

In conclusion, the AD3D-MIL algorithm not only projects the semantic representation of bags into two-class Bernoulli distribution but also can project them into the joint Bernoulli distribution of multiple classes.

E. Advantages of AD3D-MIL

1) Scalability: Intuitively, the deep instance generation module allows multiple types of data, such as text, image, video. The size of the last layer of this module can be modified according to the needed size of instances. When a new task occurs, the user only changes the size of the last layer to avoid manual pre-define and pre-process of instances. The used attention-based MIL pooling can allocate distinct weights over instances within a bag, and it allows AD3D-MIL to find multiple critical instances rather than one key instance. Moreover, the attention-based MIL pooling is trainable and fully differentiable. Finally, the last transformation function can project bag representations into the Bernoulli distributions of binary or multiple classes. These advances together can be transformed into an end-to-end neural network, and another state of the art approaches can replace each of them. Therefore, the proposed AD3D-MIL algorithm has excellent flexibility and scalability.

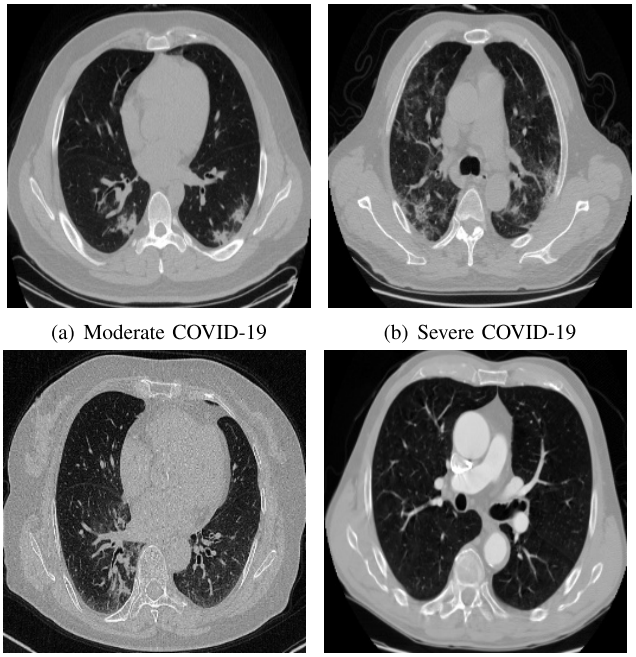
2) Interpretability: In the new task of COVID-19 screening, it is beneficial to provide infection areas together with the last screening result to the radiologists. Fortunately, the setting of multiple instance learning makes the AD3D-MIL algorithm more interpretable because the discovered key instances can indicate the location of infection areas of COVID-19. More importantly, the used attention-based MIL pooling module assigns high attention weights to instances that contribute to the positive label of the bag. It can easily interpret the provided decision and give the attention weights of instance for indicating the vital attribute of each instance. Therefore, AD3D-MIL, together with the attention mechanism, has the potential of great interest in practical applications.

V. Experiments

We evaluate the proposed algorithm on a newly-collected dataset against the state of the art methods. The entire code will be publicly available at https://github.com/zhyhan.

A. Data and Set-up

In this study, we collected a multi-class multi-center chest CT dataset comprised of 460 transverse-section CT examples. This dataset includes 230 CT examples from 79 patients with COVID-19, 100 CT examples from 100 patients with common pneumonia, and 130 CT examples from 130 people without pneumonia. The randomly selected CT images are illustrated in Figure 2. The chest CT examples from the same patient have at least tow days gap. The splitting of the training set and testing set is according to the patient-level, *i.e.,* no chest CT from the same patient exists in training and testing sets, simultaneously. Without loss of generality, the common pneumonia patients are with viral pneumonia or bacterial pneumonia. Note that these 130 people without pneumonia are either healthy or have other diseases. The dataset is collected from the designated COVID-19 hospitals in Shandong Province. Every COVID-19 patient was confirmed with nucleic acid detection kits of reverse transcription-polymerase chain reaction. The chest CT scans of COVID-19 patients without image manifestations were excluded. Moreover, the chest CT scans of common pneumonia patients are collected because it is tough yet critical to distinguish them from suspected patients with COVID-19 in clinical worldwide. This study and all research were approved and conducted following relevant guidelines/regulations.

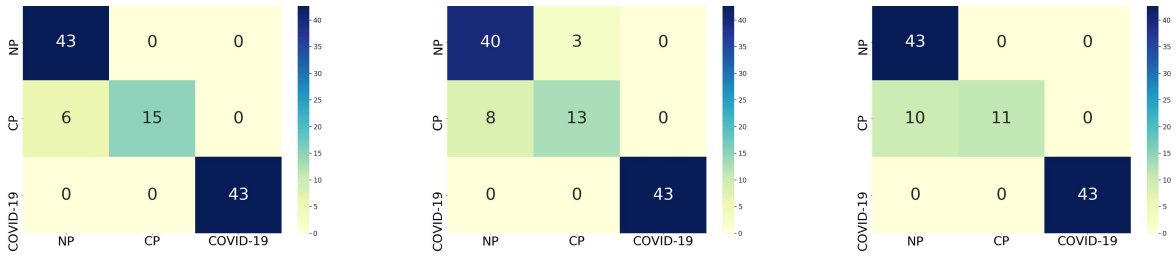


(a) Moderate COVID-19 (b) Severe COVID-19

(c) Common Pneumonia (d) Without Pneumonia

Fig. 2. The visualization of typical transverse-section chest CT slices from the collected dataset.

We conduct two screening tasks for better verifying the proposed AD3D-MIL algorithm in the problem of COVID-19 screening. The first task is the screening of COVID-19 CT scans: the positive class is COVID-19, and the negative class is Non-COVID-19. From the practical point of view, the Non-COVID-19 CT scans involve both common pneumonia and no pneumonia. The second task is the classification task of three classes: COVID-19, common pneumonia, and no pneumonia. 60% of data is used for training, 20% of data is used for model selection and super-parameters adjustment, and the remaining 20% of data is used for testing. We employ standard five-fold cross-validation on the training and validation set for adjusting super-parameters. Each experiment is repeated five times to obtain fair comparisons. The evaluation metrics include accuracy, F1 score, precision, recall, Cohen kappa



(a) AD3D-MIL (b) C3D [51] (c) DeCoVNet [11]

Fig. 3. The confusion matrix of three classes classification: COVID-19, common pneumonia (CP), and no pneumonia (NP).

TABLE I
CLASSIFICATION RESULTS ON THE BINARY CLASSES: COVID-19 AND NON-COVID-19 (COMMON PNEUMONIA, NO PNEUMONIA). STANDARD DEVIATION VALUES ARE ZEROS

Metric	Method		
	C3D [51]	DeCoVNet [11]	AD3D-MIL
Accuracy	96.8	96.8	**97.9**
AUC	98.2	98.2	**99.0**
F1 score	96.8	96.8	**97.9**
Precision	96.8	96.8	**97.9**
Recall	96.8	96.8	**97.9**
Cohen kappa score	93.6	93.6	**95.7**

score, the receiver operating characteristic (ROC) curve, and the area under the ROC curve (AUC).

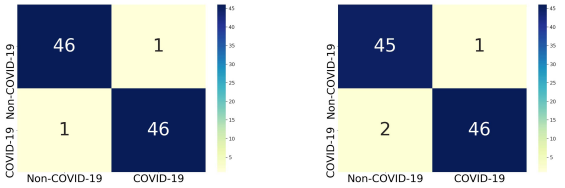
We compare our designed AD3D-MIL algorithm with state of the art methods: C3D [51] and DeCoVNet [11]. C3D is an effective approach for spatiotemporal feature learning using deep 3D convolutional networks. DeCoVNet is a newly-designed 3D deep convolutional neural network to screen COVID-19 from CT scans. C3D and DeCoVNet are supervised methods under the supervised learning setting.

We implement the AD3D-MIL algorithm in Pytorch. We use the 3D convolutional layers of DeCoVNet as the deep instance generator ψ. We set the output shape $H^* \times W^* \times S^* \times D$ of ψ be $8\times8\times8\times32$ according to cross-validation. The input shapes of CT slices are 256×256, and the slice number varies. The transformation function g is a 2-layer neural network. We set the training epoch T to 100. Data augmentation strategies, including color jittering and random affine transformation, were used. Adam optimizer is used with default parameters and an initial learning rate of 1e-5. All the compared models are implemented according to their open-source codes in Pytorch.

B. Results

1) Binary Classification: Table I reports the results on the screening of COVID-19 from chest CT. All the used algorithms are achieving promising performance. Among them, AD3D-MIL significantly outperforms the C3D and DeCoVNet models on almost all metrics. Note that while all the methods achieve promising performance, our algorithm can obtain a more interpretable result, as illustrated in Figure 8.

Figure 4 shows the confusion matrixes of AD3D-MIL, DecovNet, and C3D. The AD3D-MIL algorithm obtains a balance performance. Figure 5 illustrates the ROC curve of the AD3D-MIL algorithm, which characterizes the robustness



(a) ours (b) C3D [51] and DeCoVNet [11]

Fig. 4. The confusion matrix of the binary classification task.

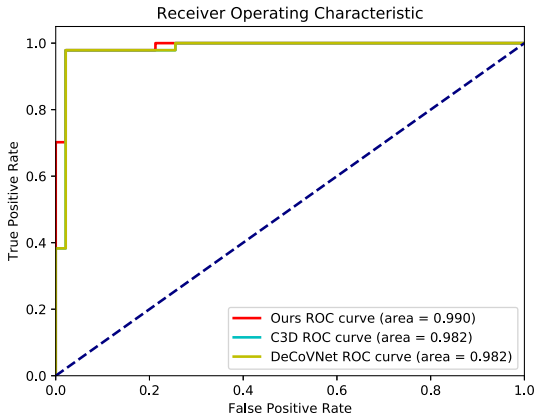


Fig. 5. The receiver operating characteristic curve of binary classification between COVID-19 and Non-COVID-19.

TABLE II
CLASSIFICATION RESULTS ON THREE CLASSES: COVID-19, COMMON PNEUMONIA, AND NO PNEUMONIA

Metric	Method		
	C3D [51]	DeCoVNet [11]	AD3D-MIL
Accuracy	89.7±0.9	90.6±0.6	**94.3±0.7**
AUC	97.1±0.4	97.5±0.1	**98.8±0.2**
F1 score	86.1±0.5	86.1±0.3	**92.3±0.4**
Precision	88.2±0.2	93.7±0.5	**95.9±0.3**
Recall	85.0±0.3	84.1±0.6	**90.5±0.5**
Cohen kappa score	83.7±0.8	84.9±0.7	**91.1±0.9**

and stability on the screening of COVID-19. From another view, these results demonstrate that the characteristic features of COVID-19 on chest CT are different from Non-COVID-19. Therefore, they are easy to be distinguished by deep models.

2) Multiple Classification: Table II reports the results on the difficult three-class classification tasks. Briefly speaking, the AD3D-MIL algorithm outperforms compared algorithms by a large margin. The AD3D-MIL algorithm obtains a classification accuracy of 94.3%, which outperforms the C3D model by 4.6% and the DeCovNet by 3.7%. Even both the spatial complexity of 3D CT scans and weak labels lead

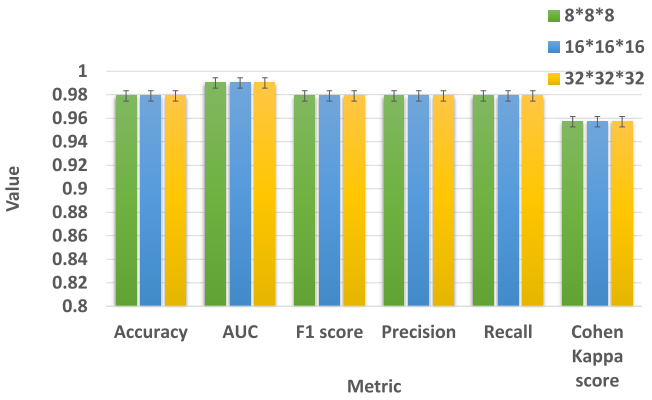


Fig. 6. Classification results on the binary classification with different instance numbers: 8*8*8 denotes that there are 512 deep instances generated from three axes of x, y, and z (best in color).

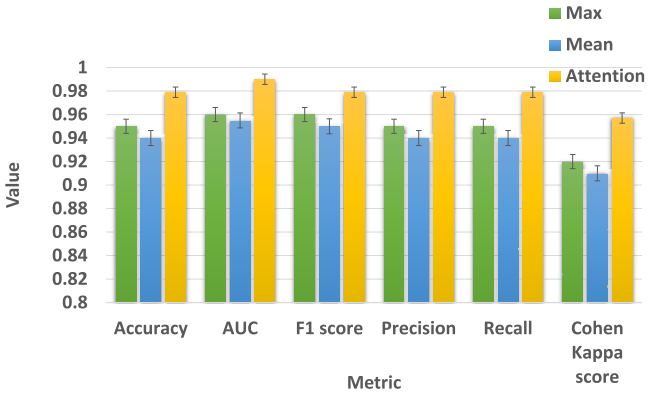


Fig. 7. Classification results on the binary classification with different instance pooling strategies: maximum, mean, attention.

to unusual difficulties, our algorithm still obtains accurate performance, which demonstrates its strengths in addressing these difficulties.

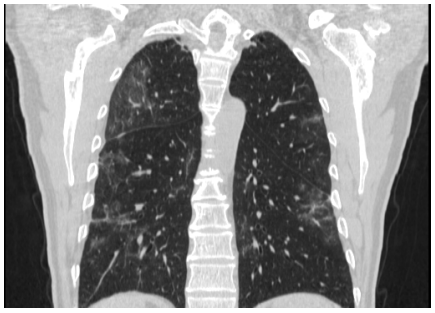
Figure 3 reports the confusion matrix to give strong evidence that AD3D-MIL obtains small predication errors and accurately screens COVID-19 without any missed case. These excellent results show that the AD3D-MIL algorithm successfully achieves accurate and robust screening of COVID-19.

We further perform statistical analysis to ensure that the experimental results have statistical significance. A paired t-test between the DeCovNet and AD3D-MIL is at a 5% significance level with a p-value of 0.008. This analysis result clearly shows that the improvement of our method is noticeable. The p-values of all compared algorithms are less than 0.05. These analyses verify that our insight that viewing the screening of COVID-19 from chest CT as a MIL problem is correct.

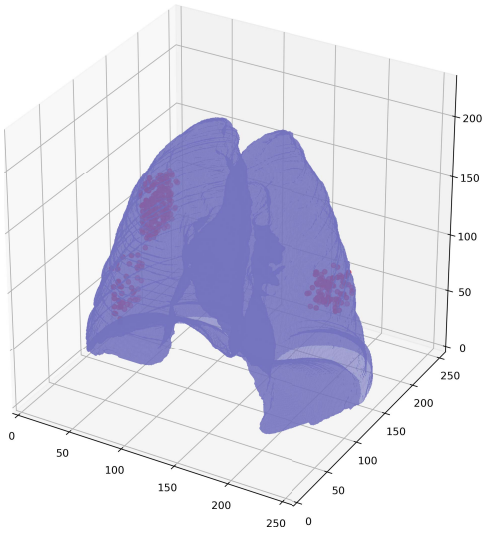
C. Analysis

This section gives an ablation study to demonstrate the effect of each new module.

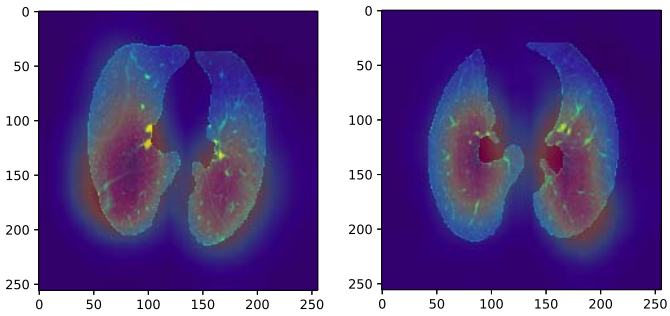
1) Number of Generated Deep Instances: Figure 6 reports the results of our ablation study on different generated instance numbers. When testing with different instance numbers, they resulted in minor changes in the proposed algorithm's performance. These results demonstrate that the flexibility and efficacy of the deep instance generator.



(a) Coronal direction CT image



(b) The discovred key instances of AD3D-MIL



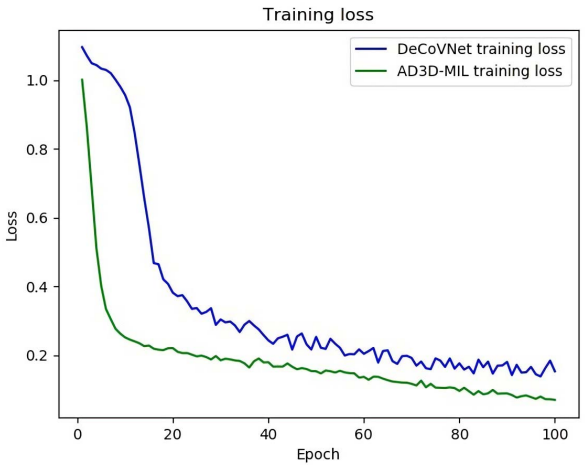
(c) DeCoVNet CAM (d) C3D CAM

Fig. 8. The visualization of key instances in COVID-19 CT. The red points in (b) indicate the infection area (best in color). (c) and (d) are class activation maps (CAM).

2) MIL Pooling Operators: We dissect the strengths of the attention-based MIL pooling. Intuitively, Figure 7 characterizes the results of our ablation study on different MIL pooling operators. The mean operator performs worse than the maximum operator. The maximum operator performs worse than attention-based MIL pooling with an extent margin. These results once demonstrate that the attention mechanism plays a crucial role in the AD3D-MIL algorithm.

3) Key Instances: While Figure 7 has demonstrated the strengths of our attention-based MIL pooling of improving the screening accuracy, we provide a broader spectrum for more in-depth analysis. Figure 8 demonstrates that the AD3D-MIL algorithm can find the key instances in accordance with the

(a) Training accuracy curve



(b) Training loss curve

Fig. 9. The statistical analysis of training and validation in the multiple classification tasks.

infected areas. While the class activation maps generated by DeCoVNet and C3D only indicate the lung areas in 2D slices coarsely, the key instances of our method precisely indicate the infection areas of COVID-19. These illustrations demonstrate the interpretability of our method. Compared to the analysis of the class activation maps, the advantages of AD3D-MIL are three-fold. Firstly, AD3D-MIL can precisely discover the infection areas of COVID-19 by key instances. Secondly, AD3D-MIL can find 3D infection areas that are more beneficial for large-scale screening of COVID-19. Class activation maps can only apply to 2D slices. Finally, the process of finding key instances is natural and easy-to-implement, while generating class activation maps is still a post-hoc analysis. We have also conducted more analyses to discover what kinds of pathology features contribute to the diagnosis of COVID-19. We mainly found that the ground-glass opacities mostly appear in the early stage and pulmonary consolidation in the late stage, which are consistent with clinical findings.

4) Training Stability: Although all the results have verified the advantages of the AD3D-MIL algorithm, we should prove its convergence and stability. Figure 9 presents the training loss and accuracy curve of AD3D-MIL and DeCoVNet on the multi-class classification task. Our newly-proposed algorithm maintains fast convergence and stable accuracy, which are more optimal than DeCoVNet's.

VI. Conclusion

We reported a new attempt of weakly-supervised screening of COVID-19 from chest CT, an under-explored but more realistic scenario. We proposed a novel attention-based deep 3D multiple instance learning (AD3D-MIL) for the screening of COVID-19 with weak labels yet high interpretability. AD3D-MIL includes a deep instance generator to generate deep 3D instances automatically, an attention-based MIL pooling to combine deep instances into an informative bag representation, and a transformation function to transform the bag representation into Bernoulli distribution or joint distributions for multiple classes of bags. The combination of these three functions can boost the generalization and interpretability of screening algorithms. Comprehensive results have demonstrated that AD3D-MIL can achieve high yet interpretable results. In-depth analyses have revealed the effectiveness and potential of AD3D-MIL as a clinical tool to relieve radiologists from laborious workloads, such that contribute to the large-scale screening of COVID-19.

References

[1] 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.

[2] M. Chung *et al.*, "CT imaging features of 2019 novel coronavirus (2019-nCoV)," *Radiology*, vol. 295, no. 1, pp. 202–207, 2020.

[3] S. Wang *et al.* (2020). A Deep Learning Algorithm Using CT Images to Screen for Corona Virus Disease (COVID-19). medRxiv. [Online]. Available: https://www.medrxiv.org/content/early/2020/04/24/2020.02.14.20023028

[4] X. Xu *et al.*, "Deep learning system to screen coronavirus disease 2019 pneumonia," 2020, *arXiv:2002.09334*. [Online]. Available: http://arxiv.org/abs/2002.09334

[5] F. Shi *et al.*, "Large-scale screening of COVID-19 from community acquired pneumonia using infection size-aware classification," 2020, *arXiv:2003.09860*. [Online]. Available: http://arxiv.org/abs/2003.09860

[6] S. Jin *et al.* (2020). Ai-Assisted CT Imaging Analysis for COVID-19 Screening: Building and Deploying a Medical AI System in Four Weeks. medRxiv. [Online]. Available: https://www.medrxiv.org/content/early/2020/03/23/2020.03.19.20039354

[7] Y. Song *et al.* (2020). Deep Learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) With CT Images. medRxiv. [Online]. Available: https://www.medrxiv.org/content/early/2020/02/25/2020.02.23.20026930

[8] 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*. [Online]. Available: http://arxiv.org/abs/2003.05037

[9] O. Gozes, M. Frid-Adar, N. Sagie, H. Zhang, W. Ji, and H. Greenspan, "Coronavirus detection and analysis on chest CT with deep learning," 2020, *arXiv:2004.02640*. [Online]. Available: http://arxiv.org/abs/2004.02640

[10] C. Jin *et al.* (2020). Development and Evaluation of an AI System for COVID-19 Diagnosis. medRxiv. [Online]. Available: https://www.medrxiv.org/content/early/2020/03/27/2020.03.20.20039834

[11] C. Zheng *et al.* (2020). Deep Learning-Based Detection for COVID-19 From Chest CT Using Weak Label. medRxiv. [Online]. Available: https://www.medrxiv.org/content/early/2020/03/26/2020.03.12.20027185

[12] J. Chen *et al.* (2020). Deep Learning-Based Model for Detecting 2019 Novel Coronavirus Pneumonia on High-Resolution Computed Tomography: A Prospective Study. medRxiv. [Online]. Available: https://www.medrxiv.org/content/early/2020/03/01/2020.02.25.20021568

[13] F. Shan *et al.*, "Lung infection quantification of COVID-19 in CT images with deep learning," 2020, *arXiv:2003.04655*. [Online]. Available: http://arxiv.org/abs/2003.04655

[14] L. Huang *et al.*, "Serial quantitative chest CT assessment of COVID-19: Deep-learning approach," *Radiology: Cardiothoracic Imag.*, vol. 2, no. 2, Apr. 2020, Art. no. e200075.

[15] F. Shi *et al.*, "Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19," *IEEE Rev. Biomed. Eng.*, early access, Apr. 16, 2020, doi: 10.1109/RBME.2020.2987975.

[16] B. Ghoshal and A. Tucker, "Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection," 2020, *arXiv:2003.10769*. [Online]. Available: http://arxiv.org/abs/2003.10769

[17] J. Zhang, Y. Xie, Y. Li, C. Shen, and Y. Xia, "COVID-19 screening on chest X-ray images using deep learning based anomaly detection," 2020, *arXiv:2003.12338*. [Online]. Available: http://arxiv.org/abs/2003.12338

[18] A. Narin, C. Kaya, and Z. Pamuk, "Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks," 2020, *arXiv:2003.10849*. [Online]. Available: http://arxiv.org/abs/2003.10849

[19] L. Wang and A. Wong, "COVID-net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images," 2020, *arXiv:2003.09871*. [Online]. Available: http://arxiv.org/abs/2003.09871

[20] T. Li, Z. Han, B. Wei, Y. Zheng, Y. Hong, and J. Cong, "Robust screening of COVID-19 from chest X-ray via discriminative cost-sensitive learning," 2020, *arXiv:2004.12592*. [Online]. Available: http://arxiv.org/abs/2004.12592

[21] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, Jan. 2018. [Online]. Available: https://academic.oup.com/nsr/article/5/1/44/4093912

[22] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, nos. 1–2, pp. 31–71, Jan. 1997.

[23] M.-A. Carbonneau, V. Cheplygina, E. Granger, and G. Gagnon, "Multiple instance learning: A survey of problem characteristics and applications," *Pattern Recognit.*, vol. 77, pp. 329–353, May 2018.

[24] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 577–584.

[25] K. Ali and K. Saenko, "Confidence-rated multiple instance boosting for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1417–1424.

[26] Q. Zhang and S. A. Goldman, "EM-DD: An improved multiple-instance learning technique," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 1073–1080.

[27] W.-J. Li and D. Y. Yeung, "MILD: Multiple-instance learning via disambiguation," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 1, pp. 76–89, Jan. 2010.

[28] Y. Xu, J.-Y. Zhu, E. I.-C. Chang, M. Lai, and Z. Tu, "Weakly supervised histopathology cancer image segmentation and classification," *Med. Image Anal.*, vol. 18, no. 3, pp. 591–604, Apr. 2014.

[29] G. Quellec *et al.*, "A multiple-instance learning framework for diabetic retinopathy screening," *Med. Image Anal.*, vol. 16, no. 6, pp. 1228–1240, Aug. 2012.

[30] T. Tong, R. Wolz, Q. Gao, R. Guerrero, J. V. Hajnal, and D. Rueckert, "Multiple instance learning for classification of dementia in brain MRI," *Med. Image Anal.*, vol. 18, no. 5, pp. 808–818, Jul. 2014.

[31] J. Melendez *et al.*, "A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest X-rays," *IEEE Trans. Med. Imag.*, vol. 34, no. 1, pp. 179–192, Jan. 2015.

[32] V. Cheplygina, L. Sorensen, D. M. J. Tax, J. H. Pedersen, M. Loog, and M. D. Bruijne, "Classification of COPD with multiple instance learning," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 1508–1513.

[33] Y. Chen and J. Z. Wang, "Image categorization by learning and reasoning with regions," *J. Mach. Learn. Res.*, vol. 5, pp. 913–939, Aug. 2004.

[34] Q. Zhang, S. A. Goldman, W. Yu, and J. E. Fritts, "Content-based image retrieval using multiple-instance learning," in *Proc. ICML*, vol. 1. New York, NY, USA: Citeseer, Jul. 2002, pp. 682–689.

[35] J. Tang, H. Li, G.-J. Qi, and T.-S. Chua, "Image annotation by graph-based inference with integrated Multiple/Single instance representations," *IEEE Trans. Multimedia*, vol. 12, no. 2, pp. 131–141, Feb. 2010.

[36] M. Dundar, B. Krishnapuram, R. Rao, and G. M. Fung, "Multiple instance learning for computer aided diagnosis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 425–432.

[37] B. Settles, M. Craven, and S. Ray, "Multiple-instance active learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1289–1296.

[38] Z. Jorgensen, Y. Zhou, and M. Inge, "A multiple instance learning strategy for combating good word attacks on spam filters," *J. Mach. Learn. Res.*, vol. 9, pp. 1115–1146, Jun. 2008.

[39] J.-Y. Zhu, J. Wu, Y. Xu, E. Chang, and Z. Tu, "Unsupervised object class discovery via saliency-guided multiple class learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 862–875, Apr. 2015.

[40] B. Babenko, M.-H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1619–1632, Aug. 2011.

[41] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," 2018, *arXiv:1802.04712*. [Online]. Available: http://arxiv.org/abs/1802.04712

[42] S. Wang *et al.*, "Computer-aided endoscopic diagnosis without human-specific labeling," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 11, pp. 2347–2358, Nov. 2016.

[43] V. Cheplygina, M. de Bruijne, and J. P. W. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Med. Image Anal.*, vol. 54, pp. 280–296, May 2019.

[44] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[45] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[46] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[47] Z. Lin *et al.*, "A structured self-attentive sentence embedding," 2017, *arXiv:1703.03130*. [Online]. Available: http://arxiv.org/abs/1703.03130

[48] N. Pappas and A. Popescu-Belis, "Explicit document modeling through weighted multiple-instance learning," *J. Artif. Intell. Res.*, vol. 58, pp. 591–626, Mar. 2017.

[49] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 652–660.

[50] J. Feng and Z.-H. Zhou, "Deep MIML network," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1884–1890.

[51] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.