Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Knowledge-Based Systems 252 (2022) 109278

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Explainable multi-instance and multi-task learning for COVID-19 diagnosis and lesion segmentation in CT images☆

Minglei Li [a], Xiang Li [a], Yuchen Jiang [a], Jiusi Zhang [a], Hao Luo [a,*], Shen Yin [b,*]

[a] Department of Control Science and Engineering, Harbin Institute of Technology, Harbin, 150001, Heilongjiang, China
[b] Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology, Trondheim, 7034, Norway

ARTICLE INFO

Article history:
Received 24 January 2022
Received in revised form 12 June 2022
Accepted 13 June 2022
Available online 27 June 2022

Keywords:
Automated diagnosis
Lesion segmentation
COVID-19
Adaptive multi-task learning
Explainable multi-instance learning

ABSTRACT

Coronavirus Disease 2019 (COVID-19) still presents a pandemic trend globally. Detecting infected individuals and analyzing their status can provide patients with proper healthcare while protecting the normal population. Chest CT (computed tomography) is an effective tool for screening of COVID-19. It displays detailed pathology-related information. To achieve automated COVID-19 diagnosis and lung CT image segmentation, convolutional neural networks (CNNs) have become mainstream methods. However, most of the previous works consider automated diagnosis and image segmentation as two independent tasks, in which some focus on lung fields segmentation and the others focus on single-lesion segmentation. Moreover, lack of clinical explainability is a common problem for CNN-based methods. In such context, we develop a multi-task learning framework in which the diagnosis of COVID-19 and multi-lesion recognition (segmentation of CT images) are achieved simultaneously. The core of the proposed framework is an explainable multi-instance multi-task network. The network learns task-related features adaptively with learnable weights, and gives explicable diagnosis results by suggesting local CT images with lesions as additional evidence. Then, severity assessment of COVID-19 and lesion quantification are performed to analyze patient status. Extensive experimental results on real-world datasets show that the proposed framework outperforms all the compared approaches for COVID-19 diagnosis and multi-lesion segmentation.

1. Introduction

Severe acute respiratory syndrome coronavirus (SARS-CoV) [1], middle east respiratory syndrome coronavirus (MERS-CoV) [2], and novel coronavirus (SARS-CoV-2) are highly pathogenic coronaviruses known to infect humans, whose infection can cause severe respiratory syndromes. At the beginning of 2020, SARS-CoV-2 quickly became a pandemic worldwide. SARS-CoV-2 infection can inflame the air sacs in the lungs and cause pleural effusion, which can lead to breathing difficulties, fever, cough, or other flu-like symptoms in patients. The diseases caused by this virus are collectively referred to as Coronavirus Disease 2019 (COVID-19) [3]. It has made a huge impact on the medical systems and economic activities around the globe. Molecular diagnostic testing and imaging testing are two main detection methods for COVID-19 infection. Reverse Transcription-Polymerase Chain Reaction (RT-PCR) as a molecular diagnostic test performed in standard laboratories, it is a common standard for the detection of COVID-19. However, the current laboratory test has a long turnaround time, and RT-PCR testing for COVID-19 may be falsely negative due to specimen contamination or insufficient viral material in the specimen, which may not be sufficient to confirm infection or infection-free [4,5].

Radiology imaging procedure has a faster turnaround time than RT-PCR, which facilitates the rapid screening of suspected COVID-19 patients in the severe and complicated novel coronavirus pneumonia epidemic. As one of the most common imaging tests, computed tomography (CT) is an effective tool for screening lung lesions and a means of diagnosing COVID-19 [6]. Chest CT images provide more detailed pathological information, which can quantitatively measure lesion size and the extent of lung involvement better [7]. Due to its convenience, accuracy, high positive rate, and good reproducibility, CT images play an indispensable role in screening COVID-19 and assessing the progression of patients. However, manual screening is a time-consuming and labor-intensive task. In addition, CT images of COVID-19 and some other common types of pneumonia have similar imaging characteristics. Sometimes even the most experienced physicians have difficulty analyzing them without relying on other testing methods. Therefore, reliable computer-aided diagnosis (CAD)

☆ This work is supported by the Young Scientist Studio of Harbin Institute of Technology, China under Grant AUGA9803503221 and Interdisciplinary Research Foundation of Harbin Institute of Technology, China under Grant IR2021224.

* Corresponding authors.
E-mail addresses: hao.luo@hit.edu.cn (H. Luo), shen.yin@ntnu.no (S. Yin).

https://doi.org/10.1016/j.knosys.2022.109278
0950-7051/

methods based on CT images are needed to boost the efficiency of COVID-19 diagnosis and assessment. Developing such methods has been a research hotspot in the past year.

Because there are huge differences in the appearance, size, and location of the pneumonia lesions, it is difficult to design an appropriate CAD method to deal with the complex features of pneumonia lesions only using machine vision methods such as classic image processing techniques or traditional statistical learning methods.

Benefiting from the development of deep learning technology, the application of convolutional neural networks (CNNs) in automated COVID-19 diagnosis has become popular. For instance, there have published several recent works that focused on automated diagnosis [8–10] or lesion segmentation of COVID-19 [11–13]. However, most existing automated diagnosis methods are two-stage or depend heavily on expert knowledge and experience. And as a result, time cost and human factors will affect the efficiency and consistency of diagnosis. In another aspect, the diagnosis task of distinguishing COVID-19 from other common pneumonia with similar symptoms has received plenty of attention, but the severity of patients is rarely assessed which can be more challenging. In addition, the diagnosis results of deep learning methods also need to be explained with clinical significance. For lesion segmentation, some studies treat the segmentation task as an intermediate step in the diagnosis of COVID-19, and a detailed quantitative or qualitative analysis of segmentation results is not conducted. The segmentation of multiple lesion areas in the lung has been demanded due to clinical potential for disease progression analysis, while it is challenging as lesions are often similar and the sizes are quite small.

It is observed that most previous works consider automated diagnosis and lesion segmentation as two independent tasks or treat segmentation as a pre-step of disease diagnosis, thus ignoring the correlation between these two tasks. During the segmentation of lesion areas, rich spatial information and tissue type information can be obtained from CT images. Similarly, the diagnosis task also needs to pay attention to such information. Since diagnosis and segmentation are highly related tasks, the multi-task learning (MTL) scheme can utilize the potential representations between these two tasks to improve the performance on each task [14]. And this MTL method is also faster than the two-stage methods of first segmentation and then classification.

Based on the above analysis, this paper proposes a multi-task learning framework that combines automated diagnosis of COVID-19 and lung multi-lesion segmentation in CT images. Specifically, the regions related to pneumonia assessment are few and unevenly distributed in each CT image of the individual. To address this problem, we treat each CT image as a bag of CT slices through a weakly supervised multi-instance learning (MIL) strategy and use these instance bags as the inputs of the following model. A multi-task network is designed, including a shared encoder for feature extraction, a diagnosis branch and a segmentation branch. Here, the shared encoder learns common representations of both tasks and generates the information exchange, the diagnosis branch and the segmentation branch are used to solve the diagnosis and segmentation tasks, respectively.

In addition, based on the results of lesion segmentation, severity assessment of COVID-19 and lesion quantification are performed. And explainability of the diagnosis results can be implemented through a dedicatedly-designed Transformer MIL Pooling (TMP) layer. Extensive experiments are conducted on real-world COVID-19 datasets to verify the effectiveness of the proposed method. The contributions of this work can be summarized as follows:

1. A multi-task learning framework is proposed to perform automated diagnosis of COVID-19 and lung multi-lesion segmentation in CT images simultaneously. Both tasks can adaptively exchange unique task-related information and learn common representations to improve the performance on each task.
2. A novel explainable multi-instance learning strategy is designed, in which a TMP layer considers the expressive abilities of different instances, constructs the local to the global representations of CT images, and can endow the diagnosis results with a certain degree of explainability by suggesting which sets of instances exert similar effects on the diagnosis.
3. The explainable multi-instance multi-task network (EMTN) in the framework is flexible, which has EM-Seg (EMTN with only the segmentation branch) and EM-Cls (EMTN with only the classification branch) two variants. Either variant of EMTN can be employed to perform diagnosis of COVID-19 or lesion segmentation in CT images.

The rest of the paper is organized as follows: Section 2 introduces the related works about the automated diagnosis and segmentation of COVID-19 based on CT images. In Section 3, we explain the proposed framework in detail. Then, the datasets and experimental setup are introduced in Section 4. The extensive experiments and discussions are presented in Section 5. Finally, Section 6 concludes this paper.

2. Related work

In this section, we briefly review the deep learning based COVID-19 classification and segmentation studies, including segmentation in CT images of COVID-19 and automated diagnosis of COVID-19 from CT images.

2.1. Segmentation of CT images infected with COVID-19

Segmentation is a key task in the assessment of COVID-19. Its main goal is to identify and mark the regions of interest (ROI) such as lung, lung lobes, or lesion areas in CT images. In terms of target ROI, the segmentation methods can be divided into two categories, namely, methods for lung fields and for lung lesions.

(1) Segmentation methods for lung fields: The methods aim to separate lung fields (i.e., the whole lung or lung lobes) from background areas. This is considered a basic task of segmentation in CT Images of COVID-19. For example, Chaganti et al. [11] proposed a two-stage method for segmenting lungs and lung lobes in CT images of COVID-19 patients. In the first stage, deep reinforcement learning was used to detect lung fields. Then, a depth image-to-image network was employed to segment the lungs and lung lobes from the detected lung fields. Xie et al. [12] introduced a non-local neural network module in the CNN to capture structural relationships of different tissues and perform the segmentation of lung lobes.

(2) Segmentation methods for lung lesions: The purpose is to segment lung lesions from other tissues in CT images. Since the lesions are of a variety of shapes and textures, and the sizes are usually small. A focus on lung lesions can improve the efficiency of follow-up in patients with COVID-19. Lung lesion segmentation is widely regarded as a challenging task. Some studies performed binary segmentation of lesions, that is, predicting the masks of all types of lesions without distinction [15–17]. For instance, Abdel-Basset et al. [15] adopted a dual-path network architecture, and designed a recombination and recalibration module that exchanges feature information to improve the segmentation of infected areas. Wu et al. [16] proposed an encoder–decoder CNN

architecture with attentive feature fusion and deep supervision strategy, and obtained the locations of the infected areas. Chassagnon et al. [17] designed a CovidENet, which is an ensemble of 2D and 3D architectures based on AtlasNet [18]. CovidENet was employed to segment the lung lesions as a whole, and it can reach a segmentation level close to physicians in terms of dice coefficient and Hausdorff distance. Moreover, some scholars carried out researches on multi-lesion segmentation methods. Chaganti et al. [11] also paid attention to multi-lesion segmentation. They used a densely connected U-Net [19] to segment ground-glass opacities (GGO) and lung consolidation. Fan et al. [20] utilized a two-stage method with two CNNs for multi-lesion segmentation. The first CNN segmented the total lesions, and then based on the segmentation results of the first CNN, the second CNN was used to further segment the GGO and lung consolidation of total lesions. Similar to Fan et al. [20], Zhang et al. [13] also adopted a two-stage method to perform lesion segmentation. Total lesions and healthy lung tissues were first segmented. After that, different types of lesions were separated from total lesions, including GGO, lung fibrosis, lung consolidation, etc.

2.2. Automated diagnosis of COVID-19 from CT images

Automated diagnosis (classification) is one crucial task in COVID-19 detection. It aims to provide rapid and accurate judgments for the diagnosis of suspected COVID-19 patients. Considering the screening COVID-19 patients as the main target, the diagnosis of COVID-19 based on CT images mainly includes two categories: (1) binary classification (COVID-19 or non-COVID-19); (2) multi-class classification (COVID-19 and other types of pneumonia).

(1) Binary classification: In this category, many studies have been carried out to distinguish COVID-19 patients from non-COVID-19 patients. Li et al. [6] utilized a modified CheXNet [21] to diagnose COVID-19. The modified CheXNet was first pre-trained on a chest X-rays dataset, and then the pre-trained network was transferred to the target task by transfer learning. Wang et al. [22] utilized a 3D U-Net [23] based segmentation model to obtain lung segments in CT images, then coupled two 3D-ResNets into a classification model via a priority attention strategy, and finally predicted the type of patients through this classification model. A weakly-supervised automated diagnosis framework [24] was established. Specifically, a pre-trained U-Net was adopted to segment lung fields. Afterwards, the segmented lung fields and original CT images were as the inputs of a 3D deep neural network, which can determine whether patients are infected with COVID-19 or not. Shaban et al. [25] proposed an enhanced KNN classifier with hybrid feature selection, which selected significant features from CT images and detected COVID-19 patients based on these features. Ardakani et al. [8] extracted 1020 CT slices from CT images of 108 COVID-19 patients and 86 non-COVID-19 patients, and ten classical CNNs were employed to distinguish between COVID-19 patients and non-COVID-19 patients based on the extracted 2D slices. However, the classification results of Ardakani et al. were slice-level rather than individual-level. This may have the problem of data leakage. Bai et al. [9] segmented the lung regions in CT images and sliced them, then an EfficientNet B4 [26] was used to obtain the classification score of each 2D slice. After that, they integrated the predicted classification scores of several 2D CT slices to make final decisions at the individual level.

(2) Multi-class classification: COVID-19 has similar imaging features in CT images to other viral, bacterial, and community-acquired pneumonia, especially viral pneumonia. It is a challenging task for physicians to judge suspected patients if the pneumonia-related lesions are subtle. The classification of COVID-19 patients from other types of pneumonia patients can accelerate the screening of patients in clinical practice. Wang et al. [27] proposed a two-stage method to perform the classification of COVID-19 and other types of pneumonia. In the first stage, a 3D DenseNet121-FPN was adopted to segment lung fields in CT images, and then in the second stage, a COVID-19Net was designed to determine the label of each patient based on the patient's clinical information and segmented lung fields. Han et al. [10] proposed a deep 3D multiple instance learning method based on the attention mechanism. With a certain number of 3D patches extracted from CT images as the inputs of a 3D CNN, the patch-level features were obtained through this CNN, and an attention-based pooling layer mapped the patch-level features into embedding space. Then, the features were transformed into the Bernoulli space, which can give the probabilities of patients with COVID-19. Song et al. [28] utilized OpenCV [29] to extract 15 complete lung slices from each CT image, and a diagnosis system was proposed to identify patients with COVID-19 from other types of pneumonia based on these slices, in which the system mainly consists of a ResNet50 fused with feature pyramid networks. Similarly, Wang et al. [30] manually selected slices with lesions according to the typical characteristics of pneumonia, and separated the CT slices into lung fields and other regions. Afterwards, they used these lung fields as the inputs of a pre-trained GoogLeNet [31] for the diagnosis of patients.

3. Methodology

3.1. Method overview

We construct a framework to perform automated diagnosis of COVID-19 and multi-lesion segmentation in CT images. The proposed framework is illustrated in Fig. 1. In order to weaken the influence of different CT images qualities, several preprocessing steps are adopted, and then a bag of instances is constructed by randomly choosing a set of 2D image slices. An explainable multi-instance multi-task network (EMTN) is designed to simultaneously perform classification and segmentation tasks based on the instance bags. As shown in Fig. 1, the architecture of EMTN consists of a shared encoder and two task branches (i.e., diagnosis and segmentation). The shared encoder plays a role in feature extraction of 2D image slices in each bag of instances, where the obtained features are fed into the following diagnosis and segmentation branches. In the diagnosis branch, these extracted features are used to construct the global representations of CT images, and the diagnosis of patients is determined through a classification layer learned from these global representations. In the segmentation branch, multi-lesion segmentation is performed via a decoding architecture symmetrical to the shared encoder. In the way of multi-task learning, task-related features such as textures and shapes can be properly utilized by both tasks to improve performance. In addition, EM-Seg (EMTN with only the segmentation branch) and EM-Cls (EMTN with only the classification branch) as two variants of EMTN, they can be employed to perform segmentation or diagnosis task individually.

After classification and segmentation, lesion quantification and severity assessment of COVID-19 patients are carried out to further enrich the functionalities of the framework. Moreover, the diagnosis results can be more explainable by observing the key instances.

3.2. Network architecture

As above-emphasized, the proposed backbone network is an encoder-focused MTL model used to simultaneously perform classification of patients and segmentation of multi-lesion. As shown in Fig. 2, the forward-propagation direction of the bags or the features in the EMTN is from the shared encoder to the

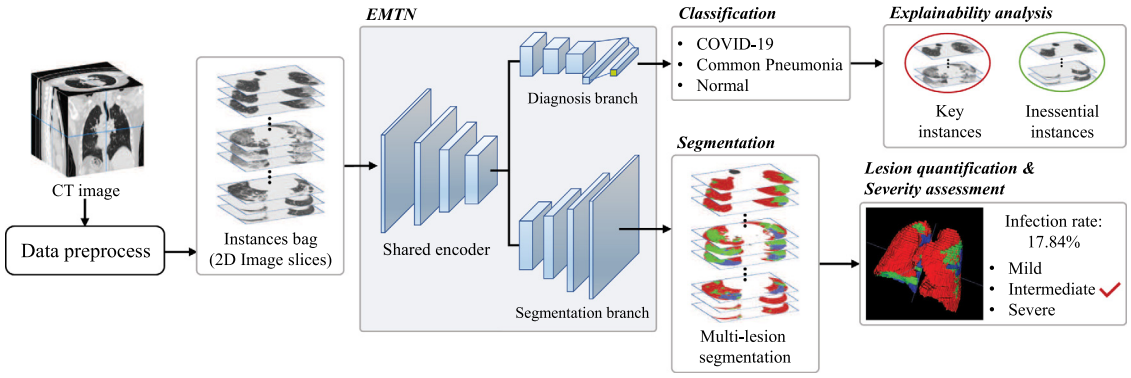

Fig. 1. Illustration of the pipeline of the proposed framework. For each CT image, a bag of instances (2D image slices) is constructed through the preprocess steps, which is fed into the explainable multi-instance multi-task network (EMTN) for the diagnosis of COVID-19 and multi-lesion segmentation. Following EMTN, explainability analysis endows the diagnosis results with explainability, lesion quantification and severity assessment provides the detailed information of patient status.

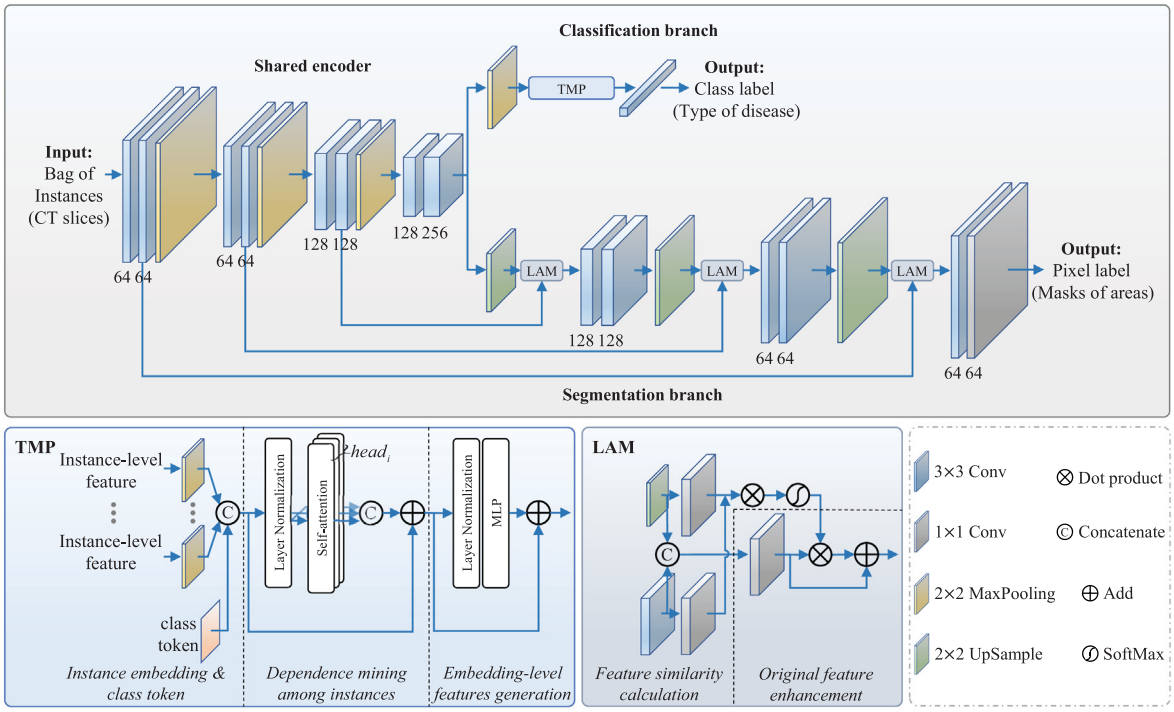

Fig. 2. The architecture of the proposed explainable multi-instance multi-task network (EMTN).

classification branch and the segmentation branch. Following this, the class label and the pixel label are determined by the corresponding branches.

Specifically, the shared encoder contains four down-sampling blocks. The first three blocks include two convolutional (Conv) layers followed by a 2×2 max pooling layer. They are designed to down-sample the intermediate feature maps. The last down-sampling block only has two Conv layers, which is designed to ensure the size of the features extracted by the encoder. The number of channels for Conv layers in four blocks are [64, 64], [64, 64], [128, 128] and [128, 256], respectively. Each Conv layer has one convolutional layer with unit stride and zero padding, followed by batch normalization (BN) and rectified linear unit (ReLU) activation. The segmentation branch has three up-sampling blocks, which forms a symmetrical codec architecture with the shared encoder. The number of channels for Conv layers in three up-sampling blocks are [128, 128], [64, 64] and [64, 64], respectively. All up-sampling blocks have the same Conv layers as down-sampling blocks, and a lesion attention module (LAM) is adopted in each up-sampling block to concatenate the up-sampled feature maps and the outputs of the corresponding down-sampling block. The LAM mainly includes feature similarity calculation and original feature enhancement two core parts. In the feature similarity calculation part, linear transformation, dot product, and feature decoupling are performed on the up-sampled feature maps and corresponding down-sampled feature maps to obtain a similarity matrix. Then, in the feature enhancement part, this similarity matrix is multiplied by the dimensionality reduced features to get the enhanced features, where the dimensionality reduced features can be obtained by concatenating and transforming the up-sampled and down-sampled features. Finally, the enhanced features are added to the original feature maps as the output of LAM. The LAM is designed to better consider the global information in the original images and enhance some tiny features beneficial to segmentation. A successful application was presented in [32]. The segmentation branch outputs the masks for the four areas (i.e., lung areas, GGO, lung consolidation, and background areas) of image slices in the bag. The classification (and

also diagnosis) branch contains a 2×2 max pooling layer, a TMP layer, and a classification layer. In this classification branch, local instance-level features are mapped to the embedding-level space through a novel explainable MIL strategy, and these embedding-level features are further processed by classification layer to obtain predicted diagnosis results.

Both tasks are simultaneously supervised by the MTL loss function which is usually a combination of single task loss functions. Since the loss functions of different tasks may have unbalanced contributions when optimizing the model parameters, a weight-adaptive multi-task learning loss function is designed in this paper. Specifically, let $\{[(X_n, Y_n), (x_{nl}, \mathcal{G}_l)]\}_{n=1}^{N}$ be a training set containing N samples, where X_n and $Y_n \in \{1, \ldots, C_{cls}\}$ denote the bag of instances for the nth CT image and the corresponding class label, respectively; x_{nl} denotes the lth instance in X_n, and $\mathcal{G}_{nl}$ is the corresponding ground-truth segmentation mask for x_{nl}. For single task, the loss functions $\mathcal{L}_{cls}$ and $\mathcal{L}_{seg}$ for classification and segmentation tasks are as follows.

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c_{cls}=1}^{C_{cls}} log(P_{cls}(\hat{Y}_n = c_{cls} \mid X_n)) \tag{1}$$

$$\begin{aligned}\mathcal{L}_{seg} = &-\frac{1}{N}\sum_{n=1}^{N}\sum_{c_{seg}=1}^{C_{seg}} log(P_{seg}(\hat{\mathcal{G}}_{nl} = c_{seg} \mid x_{nl})) \\ &- 2\sum_{c_{seg}=1}^{C_{seg}}\left(\frac{\hat{\mathcal{G}}_{nl} \cap \mathcal{G}_{nl}}{\hat{\mathcal{G}}_{nl} + \mathcal{G}_{nl}}\right) + 1\end{aligned} \tag{2}$$

where $\mathcal{L}_{cls}$ is the cross-entropy loss, and $\mathcal{L}_{seg}$ is the aggregation of cross-entropy loss and dice loss. In terms of a given bag (e.g., X_n) being diagnosed as a specific class (e.g., $\hat{Y}_n = c_{cls}$) or a given instance (e.g., x_{nl}) being segmented as a predicted mask (e.g., $\hat{\mathcal{G}}_{nl} = c_{seg}$), function $P_{cls}(\cdot|\cdot)$ and $P_{seg}(\cdot|\cdot)$ denote the probability obtained by the classification branch and segmentation branch, respectively. In this work, three classification targets and one multi-categories semantic segmentation are performed, namely C_{cls}=3, C_{seg}=4.

Then, the single task loss function is weighted by the trade-off factors λ to form a joint loss function.

$$\begin{aligned}\mathcal{L}_{joint} = &\frac{1}{2 \times \lambda_{cls}^2} \cdot \mathcal{L}_{cls} + \ln(1 + \lambda_{cls}^2) + \\ &\frac{1}{2 \times \lambda_{seg}^2} \cdot \mathcal{L}_{seg} + \ln(1 + \lambda_{seg}^2)\end{aligned} \tag{3}$$

where λ_{cls} and λ_{seg} are the trade-off factors for the classification task and segmentation task, respectively. The factors λ are learnable parameters that can be adjusted in the process of model optimization. The regularization terms [33] $R(\lambda) = \ln(1 + \lambda^2)$ are added to avoid the trade-off factor of each task reaching minimum, maximum, negative values, or zero solution.

3.3. Explainable multi-instance learning

COVID-19-related infection regions usually exist in some partial regions of the lung, while the regions in CT images are unlabeled, i.e., only the entire CT images have the corresponding labels. This situation is seen as a weakly supervised problem which can be solved with multi-instance learning (MIL) strategy. In this section, the details of the proposed explainable MIL strategy are introduced.

As mentioned in Section 3.1, a bag of image slices is constructed as the input of the network. Let $X_i = \{x_{i1}, x_{i2}, \ldots, x_{in_i}\}$ denotes the bag which represents the ith CT image, where $x_{kl} \in \mathbb{R}\ (l = 1, 2, \ldots, n_k)$ represents the lth slice of the kth CT image. And then, the shared encoder performs feature extraction on these CT slices to obtain instance-level features $S_i = \{s_{i1}, s_{i2}, \ldots, s_{in_i}\}$, followed by a proposed Transformer MIL Pooling (TMP) layer to generate embedding-level features B_i from instance-level features.

The proposed TMP layer is one of the few attempts that combine transformers [34] with MIL. Medical images usually have explicit sequences, such as modalities, slices, patches, etc. From these sequences, some important long-range dependence and semantic information can be mined. Transformers can effectively focus on long-range dependence in the data sequences. In the process of feature learning, MIL also needs to pay attention to the dependence among instances to generate embedding-level features. As a useful tool to process sequence relations, Transformers combine instance-level features with MIL, which can not only provide the degree of connection between instances and diagnosis results, but can also further improve the performance of the network. This is the motivation for us to propose TMP. Different from transformers which only analyze a single image, the TMP layer in this work is a core component of explainable MIL. It operates on features in the instance feature space and the embedding feature space, and generates embedding-level features, which can also provide the explainability of diagnosis results.

As shown in Fig. 2, the designed TMP layer contains the following three parts in series.

(1) Instance embedding and class token: For each extracted instance-level feature $s_{kl} \in \mathbb{R}^{C \times H \times W}$, C is the number of channels, (H, W) is the size of the feature map. In order to facilitate the propagation of features among different layers, we flatten the instance-level features and concatenate them into a sequence of embedded instance-level features $S_i' = [s_{i1}'; s_{i2}'; \ldots; s_{in_i}'] \in \mathbb{R}^{N \times (CHW)}$, where $s_{kl}' \in \mathbb{R}^{1 \times (CHW)}$ is the reshaped instance-level feature map, N is the number of instances, which also serves as the input sequence length for the TMP. Then, the embedded instance-level features are mapped to D dimensions with a trainable linear projection. The above process is instance embedding.

Considering that it is unfair to choose one of the embedded instance-level features as the feature for subsequent classification, a learnable vector is prepended to the sequence of embedded instance-level features, which is referred to as class token s_{class}. The class token plays the role of the bag representation, which will learn in the following steps and be used to predict the class label.

The resulting sequence of embedded instance-level features and the class token is as follows:

$$z_i = [s_{iclass}; s_{i1}'E; s_{i2}'E; \ldots; s_{in_i}'E] \in \mathbb{R}^{(N+1) \times D} \tag{4}$$

where z_i and s_{iclass} indicate the resulting sequence and the class token of ith CT image, respectively. $E \in \mathbb{R}^{(CHW) \times D}$ is the trainable linear projection.

(2) Dependence mining among instances: The core of this step is multi-head self-attention (MSA) [34] which mines the dependence in the resulting sequences based on the calculation of similarity among the elements of the sequence. MSA is an extension of self-attention (SA), which performs SA operation several times in parallel, and concatenates their outputs as the output of MSA.

For the ith element in the resulting sequence z, it is transformed into three vectors: query vector q_i, key vector k_i, and value vector v_i, where q_i and k_i are used to compute an attention weight A_{ij} that indicates the dependence with the jth element, v_i represents the unique feature map of the element. Then, a weighted sum over the values V of all elements in the sequence is computed (Eq. (5)).

$$[Q, K, V] = zW_{qkv}, \quad W_{qkv} \in \mathbb{R}^{D \times 3d} \tag{5}$$

$$A = softmax(\frac{QK^T}{\sqrt{d}}), \quad A \in \mathbb{R}^{(N+1)\times(N+1)} \tag{6}$$

$$SA(z) = AV \tag{7}$$

where Q, K and V are the sets of q, k and v, respectively. W_{qkv} is a trainable linear projection which transforms the elements in the sequence to the corresponding q, k, and v. A is the attention weight which reflects the dependence between the class token and instances. In addition, the dependence among instances can also be calculated based on A.

$$MSA(z) = [SA_1(z); SA_2(z); ...; SA_h(z)]W_{msa} \tag{8}$$

where $SA_h(z)(h = 1, 2, \ldots, h)$ represents performing h SA operations in parallel. $W_{msa} \in \mathbb{R}^{(h\times d)\times D}$ is a trainable parameter matrix that controls the output size of feature maps as D dimensions.

In this step, the MSA is with skip connections like residual networks and with a layer normalization (LN) to normalize each element in the sequence. Therefore, with the resulting sequence z in the first step as input, the second step can be described by the following equation:

$$z' = MSA(LN(z)) + z \tag{9}$$

where z' is the output of the second step.

(3) Embedding-level feature generation: In this step, the embedding-level representations are generated with a multi-layer perceptron (MLP). The MLP is composed of two linear layers with dropout and a Gaussian error linear unit (GELU) activation which is inserted between these two layers. Like MSA in the second step, the MLP also has the residual connection and a layer normalization. This step can be written as follows.

$$z'' = MLP(LN(z')) + z' \tag{10}$$

where z'' is the output embedding-level features. As mentioned in step 1, a class token is prepended to the sequence as the subsequently used bag representation. After three steps of learning, the learned class token $z''[0]$ is employed to predict the final classification.

The above three steps are the details of TMP. Furthermore, to provide the explainability of diagnosis results, the proposed explainable MIL strategy can not only process the weakly supervised problem but also suggest key instances which exert similar effects on the diagnosis. The acquisition process of the key instances is described as follows.

At each SA operation, we get an attention matrix A_h that defines the attention maps from the output class token to the input instance-level space. To process the attention matrices of multiple SA operations, we take the average of these matrices across all SA operations and normalize them to get the averaged attention matrix A_{att}. Then, the attention weight between the class token and each instance-level feature is obtained with A_{att}. Specifically, the class token corresponds to bag representation, and the predicted result is obtained by mapping the bag representation to the label space. Thus, the class token can represent the bag of instances. Each instance-level feature corresponds to a unique instance in the bag (a 2D image slice). It is straightforward that the higher attention weight between the class token and one instance-level feature, the greater impact this instance-level feature has on the bag representation and the final diagnosis result. The key instances can be selected according to the values of the attention weights, and these instances usually contain more lungs or lesion areas than other instances, which provides the basis for the posterior test of the diagnosis results.

This explainable MIL is different from the standard multi-instance assumption. The standard multi-instance assumption states that a bag is positive if and only if it contains at least one positive instance, which relies heavily on the correctness of labels and is susceptible to false positive instances. The explainable MIL strategy considers the relationship between instances and bag representation, and assigns suitable weights for instance-level features to generate embedding-level features and bag representation. This allows the generated bag representation can better characterize the samples and reduce the interference of false positive instances.

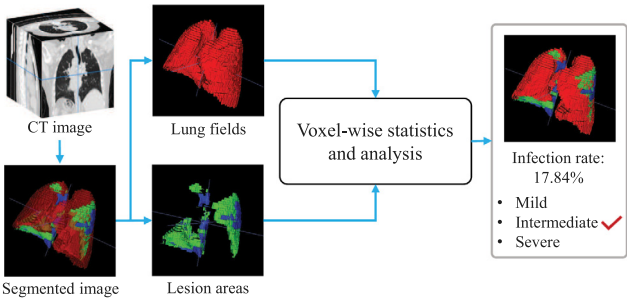

Fig. 3. The pipeline of the lesion quantification and severity assessment via voxel-wise analysis.

3.4. Quantitative analysis of lesions and severity assessment of COVID-19

The quantitative analysis of lesions and severity assessment of COVID-19 can provide more detailed information for the evaluation of patient status, clinical treatment effect, or drug experiments. In this section, we use a voxel-wise analysis method to further analyze the segmentation results.

As shown in Fig. 3, at the segmentation branch in EMTN, the positions and sizes of the lung fields, as well as GGO and lung consolidation in CT images can be obtained. In order to facilitate the analysis, GGO and lung consolidation are collectively referred to as the lesion areas. Briefly, we count the number of voxels in lung fields and lesion areas. Let n_{LF}, n_{GGO} and n_{CO} denote the number of voxels in the lung fields, GGO and lung consolidation, respectively. Then, the percentage of the number of voxels in the lesion areas to that in the lung fields is calculated as follows.

$$r = \frac{n_{GGO} + n_{CO}}{n_{LF}} \times 100\% \tag{11}$$

where r is the key indicator for quantitative analysis of lesions, i.e., infection rate. Furthermore, the severity of COVID-19 patients can be assessed with r.

4. Experimental setup

4.1. Dataset and preprocess

The real-world datasets of CT images are acquired from three sources: China Consortium of Chest CT Image Investigation (CC-CCII),[1] COVID19-CT-dataset,[2] and Radiopaedia.org.[3] The combined dataset contains CT images of COVID-19 patients, common pneumonia patients, and normal control cases in three groups. Fig. 4 shows the axial surface of CT images in different groups.

The CC-CCII dataset consists of a total of 2778+ CT images from 2778 patients [13]. These patients and their CT images are

[1] http://ncov-ai.big.ac.cn/download?lang=en

[2] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/6ACUZJ

[3] https://radiopaedia.org/cases?lang=us

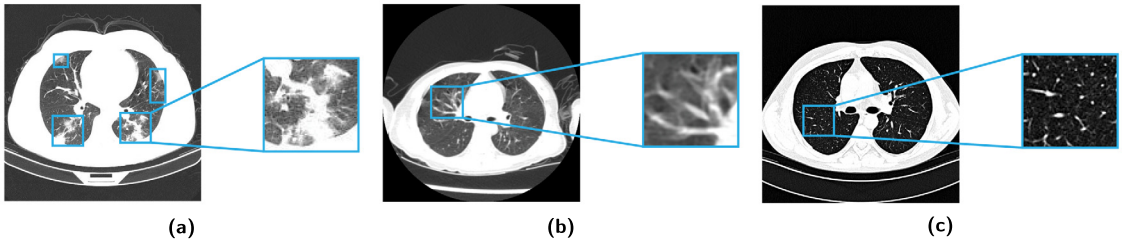

Fig. 4. Axial surface of CT images and variations in CT images of different groups. (a) One case of COVID-19 group. The COVID-19-specific patterns of lung infections are denoted by blue boxes, including ground-glass opacities and consolidation. (b) One case of common pneumonia group. The increased and/or disordered lung marking is (denoted by blue boxes) the CT manifestations of early common pneumonia patients. (c) One case of normal control group. The black area is normal lung tissue, and the white dots are normal physiological structures such as blood vessels, bronchi, etc.

Algorithm 1: Automated diagnosis and lung lesion segmentation of COVID-19

Input: The real-world CT images collected from patients with different types of pneumonia (mainly include COVID-19, common pneumonia, and normal control).

Output: The predicted diagnosis results of patients $\hat{Y}_n$;the lung segmentation masks of CT images $\mathcal{G}_{nl}$; the explainability analysis results by suggesting *key instances*;the infection rate r of the whole lung and the *severity level* of COVID-19 patient.

Data preprocess is performed to construct the instances bag;
Initialize variable n=1, the number of patients N in the algorithm;
while $n \leq N$ **do**
  With instances bag as input of shared encoder, obtain instance-level features;
  `// Classification`
  Explainable multi-instance learning strategy is employed to generate embedding-level features (according to Eqs. (4)–(10));
  Predict diagnosis results $\hat{Y}_n$ of patient n from the classification layer;
  `// Segmentation`
  Simultaneously predict lung segmentation masks of CT images $\mathcal{G}_{nl}$ by segmentation branch;
  `// Explainability analysis`
  Get the averaged attention matrix A_{att}, choose key instances according to the value of the elements in A_{att} and show them;
  `// Quantitative analysis and severity assessment`
  if $\hat{Y}_n$ *represents COVID-19* **then**
    Calculate infection rate r by Eq. (11);
    Obtain the severity level of COVID-19 patient according to the severity level definitions;
    return $\hat{Y}_n$, $\mathcal{G}_{nl}$, *key instances*, r, *severity level*;
  else
    return $\hat{Y}_n$, $\mathcal{G}_{nl}$, *key instances*;
  end
  $n \leftarrow n + 1$;
end

collected from six hospitals, including 917 COVID-19 patients, 983 common pneumonia patients, and 878 normal control cases. All COVID-19 patients are tested positive by RT-PCR. The common pneumonia patients are confirmed based on standard clinical, radiological, and molecular test results. The segmentation of CT images of COVID-19 patients is manually annotated and reviewed by five experienced radiologists. The annotated range mainly contains lesion areas (i.e., GGO and lung consolidation) which are used to distinguish COVID-19 from other cases, and to distinguish lung fields from background areas in CT images. Since part of CT images masked other regions than the lungs, and this part images cannot provide complete information of patients, so these images are screened out. In addition, those CT images with a small number of slices are also screened out. After screening, the CC-CCII dataset includes 529 COVID-19 patients, 592 common pneumonia patients, and 520 normal control cases.

The COVID-19-CT-dataset is an open-access chest CT image repository, it consists of 3000+ CT images from 1013 patients with confirmed COVID-19 infections [35]. These patients and their CT images are collected from two general hospitals in Mashhad, Iran. All COVID-19 patients have positive RT-PCR tests and accompany by supporting clinical symptoms at the point of care in an in-patient setting. And CT images of patients are visually evaluated by two board-certified radiologists to confirm the presence of COVID-19 infection. For the CT images in which the first two radiologists are in disagreement, the final decision is rendered by a third more experienced radiologist.

Radiopaedia is a non-profit, international collaborative, open-edit radiology resource compiled by radiologists and other health professionals from across the globe. We collect 130 common pneumonia patients from Radiopaedia as a supplement. The collected common pneumonia patients are confirmed by their contributors based on standard clinical, radiological, or molecular test results.

The quality of CT images is inconsistent due to the use of different CT scanners or different scanning parameters during the image acquirement process, it is a challenging task to diagnose based on these images. To reduce the risk of data leakage and to ensure the quality of images in the datasets, we make sure that only one scan (CT image) per patient is selected and each CT image is checked. Then, the following processing procedures are performed.

Specifically, the axial slices are first extracted from each CT image according to the physical spacing between slices [Fig. 5(a)]. To reduce the influence of noise in the images, each axial slice is processed with morphological operations, including structuring elements acquisition, erosion, and dilation [Fig. 5(b)]. Then, the images are binarized by Otsu algorithm [Fig. 5(c)]. With the above steps, the basic contours of human body and lung fields are identified. We detect the key points of the inner and the outer contours respectively, and identify the inner contour by the size of enclosed area [Fig. 5(d)]. Morphological operations, binarization, and contours detection are only used to determine the rectangle region of interest (ROI), therefore these operations will not lead to a loss of the image information. After that, according to the shape of the inner contours, the axial slices are

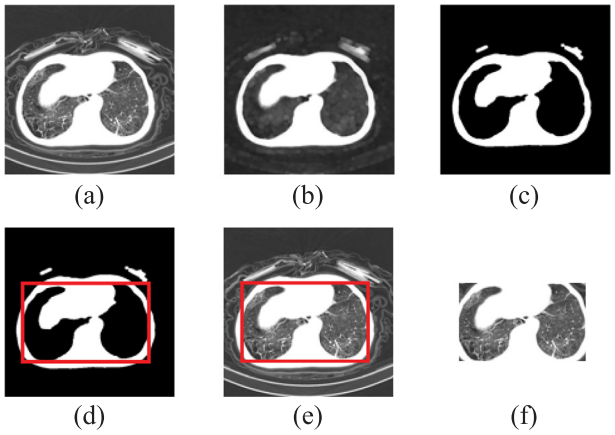

Fig. 5. Procedures of preprocessing. (a) Extract the axial slices. (b) Morphological operation on the image. (c) Binarize image by thresholding. (d) Identify the basic contour and make mask. (e) Multiply the image with mask. (f) Crop the image to get the ROI.

cropped to obtain the corresponding ROIs [Fig. 5(e–f)]. Finally, with the MIL strategy, a set of preprocessed axial slices in each CT image are randomly selected to construct a bag of instances, and each bag can represent a corresponding individual and serve as the input of EMTN. In this way, we expect that the problems of uneven data quality and data discrepancies can be alleviated.

4.2. Evaluation metrics

The proposed framework performs both classification of COVID-19 patients and segmentation of multi-lesion two tasks. In order to verify the effectiveness of the overall framework, 5-fold cross-validation is adopted, and different metrics are used in the performance evaluation.

For the classification task, five evaluation metrics are used to evaluate the classification performance, including Accuracy (ACC), Precision, Recall, F1 Score, and the area under the receiver operating characteristic (ROC) curve (AUC). The descriptions of these metrics are given below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{15}$$

where TP, TN, FP, and FN are the number of true positive samples, true negative samples, false positive samples and false negative samples, respectively. It should be noted that ROC curve and AUC only partially summarize or explain the performance of diagnosis, especially when the data is unbalanced or the values of AUC of different classifiers are close, judging a better curve is difficult. The leftmost partial area of ROC curve is the region of interest for classifying fewer positives than negatives, and ROC curves of better classifiers usually go up quickly or stay to the left side. A concordant partial AUC ($pAUC_c$) focuses on the region of interest in the ROC, and provides a good explanation for partial areas in ROC curves. The $pAUC_c$ is a foundational partial measure, which has all the interpretations offered by the AUC [36]. For an ROC curve $y = r(x)$, the definition of $pAUC_c$ is as follow:

$$pAUC_c \triangleq \frac{1}{2}\int_{x_1}^{x_2} r(x)dx + \frac{1}{2}\int_{y_1}^{y_2} 1 - r^{-1}(y)dy \tag{16}$$

where x and y are false positive rate (FPR) and true positive rate (TPR), respectively. $r^{-1}(\cdot)$ denotes inverse function. In this work, area measures are performed for three parts of an ROC curve i = 1,2,3, including the leftmost partial curve (i=1, FPR=[0.00,0.33]), the middle partial curve (i=2, FPR=[0.33,0.66]), and the rightmost partial curve (i=3, FPR=[0.66,1.0]).

For the segmentation task, three commonly used metrics are adopted to evaluate the segmentation performance, including Dice Score (DC), Positive Predict Value (PPV), and Sensitivity (SEN). The evaluation metrics are defined as:

$$DC = \frac{2|\mathcal{R}_{seg} \cap \mathcal{R}_{GT}|}{|\mathcal{R}_{seg}| + |\mathcal{R}_{GT}|} \tag{17}$$

$$PPV = \frac{|\mathcal{R}_{seg} \cap \mathcal{R}_{GT}|}{|\mathcal{R}_{seg}|} \tag{18}$$

$$SEN = \frac{|\mathcal{R}_{seg} \cap \mathcal{R}_{GT}|}{|\mathcal{R}_{GT}|} \tag{19}$$

where $\mathcal{R}_{seg}$ and $\mathcal{R}_{GT}$ denote the predicted segmentation masks and the ground-truth, respectively.

4.3. Implementation details

For the CC-CCII dataset, 70% of the samples and 20% of the samples are respectively selected as the training set and validation set to supervise the training of EMTN, and 10% of the samples are as the testing set to evaluate the performance. Since there has no ground truth for segmentation contained in the COVID-19-CT dataset and common pneumonia patients collected from Radiopaedia, both of them are used as an additional test set to evaluate the performance of EMTN, which do not participate in the training stage of EMTN. Furthermore, the samples in these two datasets can be used to train and test a variant of EMTN, i.e., EMTN with only the classification branch, and the data split also follows a 7:2:1 ratio.

The proposed method is implemented with *Python*, in which EMTN is implemented based on the *PyTorch* deep learning library. The training process is carried out on a workstation equipped with four NVIDIA GTX 1080Ti graphics cards. In both stages of training and testing, we randomly select a certain number of axial slices from each CT image to construct a bag of instances, where the bag size varies within {30, 50, 80, 100, 120, 150}, and the sizes of all bags constructed from different CT images are the same. During the training process, the EMTN is optimized by Stochastic Gradient Descent (SGD) algorithm with 100 epochs, and the weight decay is 1×10^{-4}. The initial learning rate is 0.0001, and it decays by 35% every 20 epochs.

5. Experiments and analysis

In this section, comparison study (Section 5.1) examines the superiority of the proposed method, ablation study (Section 5.2) evaluates the effectiveness of the MIL and MTL strategies. Both comparison study and ablation study are performed based on the CC-CCII dataset. Additional study (Section 5.3) evaluates the performance of the proposed method on the COVID-19-CT dataset and common pneumonia patients collected from Radiopaedia. Section 5.4 discusses the limitations of this work and future research direction.

5.1. Comparison study

We first test the performance of the proposed EMTN on both classification and segmentation tasks. Since the datasets in most of the methods listed in Section 2 are not accessible, the comparison of these methods is not possible. To show the superiority of the EMTN, it is compared with several state-of-the-art

Table 1
Comparison of classification results of different methods on three targets (bag size=100).

Targets	Method	ACC(%)	AUC(%)
COVID-19 *vs*. NC	ResNet18-Voting	88.97 [85.52, 92.37]	90.08 [86.30, 92.18]
	Gated-Attention	94.48 [92.07, 97.24]	94.63 [91.65, 97.08]
	EM-Cls	96.73 [95.00, 98.80]	96.98 [93.80, 98.99]
	EMTN	98.62 [97.59, 100.0]	98.90 [97.65, 100.0]
COVID-19 *vs*. CP	ResNet18-Voting	81.28 [76.68, 85.74]	81.63 [78.03, 84.08]
	Gated-Attention	90.34 [87.19, 93.85]	90.68 [86.95, 93.15]
	EM-Cls	93.10 [90.33, 95.89]	92.53 [89.83, 95.72]
	EMTN	95.17 [92.28, 97.38]	95.87 [91.49, 98.65]
CP *vs*. NC	ResNet18-Voting	84.93 [80.86, 89.15]	85.72 [81.38, 88.65]
	Gated-Attention	91.03 [87.99, 94.43]	91.00 [87.19, 94.22]
	EM-Cls	93.79 [91.15, 96.45]	94.03 [90.39, 97.12]
	EMTN	95.86 [93.95, 98.23]	96.46 [92.97, 98.79]

The upper and lower bounds of 95% confidence interval are shown in [·].

Table 2
Comparison of area measure $pAUC_c$ of different methods on three targets (bag size=100).

Targets	Method	$pAUC_c$(%)		
		i = 1	i = 2	i = 3
COVID-19 *vs*. NC	ResNet18-Voting	53.04	20.37	16.67
	Gated-Attention	60.96	17.00	16.67
	EM-Cls	63.64	16.67	16.67
	EMTN	65.56	16.67	16.67
COVID-19 *vs*. CP	ResNet18-Voting	46.96	17.50	17.17
	Gated-Attention	56.00	18.01	16.67
	EM-Cls	58.69	17.17	16.67
	EMTN	62.53	16.67	16.67
CP *vs*. NC	ResNet18-Voting	50.03	19.02	16.67
	Gated-Attention	54.97	19.36	16.67
	EM-Cls	59.35	18.01	16.67
	EMTN	63.12	16.67	16.67

Table 3
Comparison of multi-lesion segmentation results of different methods.

Method	DC(%)	PPV(%)	SEN(%)
U-Net	95.83 ± 3.1	96.00 ± 3.5	95.67 ± 2.7
U-Net++	96.10 ± 2.3	96.16 ± 2.8	96.05 ± 1.9
EM-Seg	94.29 ± 3.4	94.49 ± 5.0	94.08 ± 3.9
EMTN	96.18 ± 3.7	96.26 ± 4.2	96.09 ± 3.8

The results are shown as mean ± standard deviation.

classification and segmentation methods. The competing classification methods in [37,38] are the non-MIL method and the normal MIL method, respectively. ResNet18-Voting method [37] is one of the common ensemble methods and is often used in the field of patch-level or slice-level medical image analysis. It uses ResNet18 as the backbone network and performs individual-level classification through voting. Gated attention MIL method [38] combines attention mechanism with MIL to replace the pooling operators. Furthermore, two state-of-the-art segmentation methods U-Net [19] and U-Net++ [39] are compared for multi-lesion segmentation. U-Net [19] designs the commonly used symmetric U-shaped architecture, which is a classic image segmentation method. Based on U-Net, U-Net++ [39] adopts dense skip-connections to improve the fluidity of gradient, which connects the semantic gap between feature maps in the compression path and the expansion path. These methods are applied to the datasets used in this work, whose main architectures and parameter settings are consistent with those in their respective papers.

5.1.1. Classification of COVID-19

We compare and analyze the performance on three different classification targets of EMTN, EM-Cls, ResNet18-Voting, and Gated-Attention. EM-Cls represents a variant of EMTN, i.e., EMTN with only the classification branch. The classification targets include (1) COVID-19 patients vs. normal control cases (denoted as COVID-19 *vs*. NC), (2) COVID-19 patients vs. common pneumonia patients (denoted as COVID-19 *vs*. CP), and (3) common pneumonia patients vs. normal control cases (denoted as CP *vs*. NC). Table 1 shows the comparison on ACC and AUC.

It can be observed that three MIL methods (i.e., EMTN, EM-Cls, and Gated-Attention) have satisfactory performance on different classification targets, which are better than the non-MIL method ResNet18-Voting. Taking the COVID-19 *vs*. CP target as an example, the ACC achieved by EMTN, EM-Cls and Gated-Attention are 17.09%, 14.54% and 11.15% higher than ResNet18-Voting, respectively, and AUC is 17.44%, 13.35% and 11.09% higher than ResNet18-Voting, respectively. In the MIL methods, the proposed EMTN achieves the best results on the three classification targets, which can reach 98.62% ACC and 98.90% AUC on COVID-19 *vs*. NC target, 95.17% ACC and 95.87% AUC on COVID-19 *vs*. CP target, 95.86% ACC and 96.46% AUC on CP *vs*. NC target. The ROC curves of these methods are illustrated in Fig. 6, and the $pAUC_c$ of the corresponding methods are shown in Table 2 to explain the partial areas in ROC curves. It can be learned that the ROC curve of EMTN is better than others. And the EMTN achieves the best $pAUC_c$ in the leftmost partial curve area, which means the ROC curves of EMTN rise faster and end with higher TPR than other curves. Table 1, Table 2, and Fig. 6 indicate that the proposed EMTN yields superior performance in these classification targets based on CT images.

5.1.2. Segmentation of multi-lesion

We then compare and analyze the performance of EMTN, EM-Seg, U-Net, and U-Net++ on multi-lesion segmentation. EM-Seg represents that EMTN with only the segmentation branch. Table 3 shows the segmentation results of different methods. The proposed EMTN achieves the best results, with the DC of 96.18%, the PPV of 96.26%, and the SEN of 96.09%. The other methods have similar results as EMTN in a part of evaluation metrics. For example, U-Net++ has the DC of 96.10%, and the SEN of 96.05%, which are only slightly lower than ours; U-Net yields the PPV of 96.00%, while the PPV in our method is 96.26%. Compared with other methods, the performance of EM-Seg in terms of evaluation metrics is decreased, but considering that the parameter number of EM-Seg is only 1.87M, the segmentation performance is acceptable. In addition, the segmentation branch in EMTN is based on EM-Seg, and with multi-task learning, EMTN can also have a better performance.

The visualization of segmentation results achieved by four different methods is shown in Fig. 7. It can be seen that the segmentation masks generated by EMTN are more complete and have fewer missing segmentation results than other methods. In

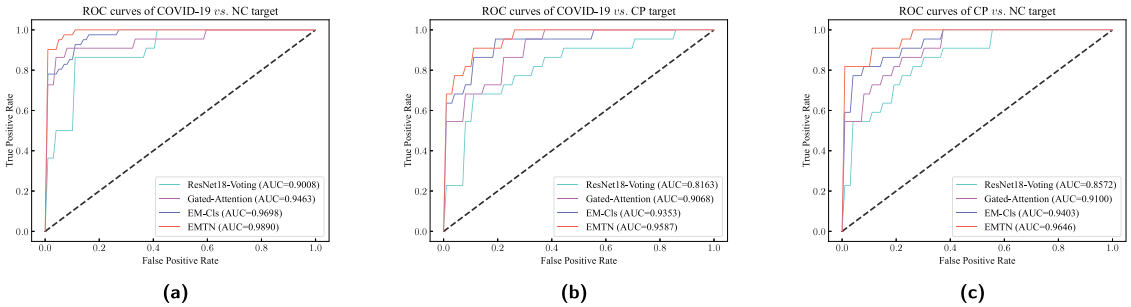

Fig. 6. ROC curves of different classification targets achieved by four methods.

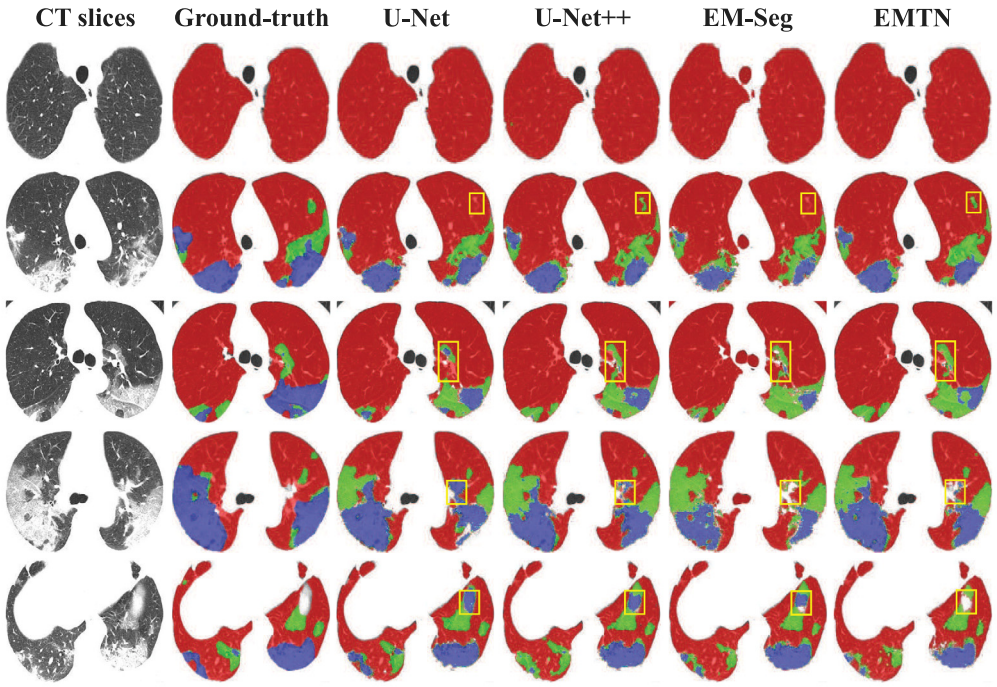

Fig. 7. Visualization of multi-lesion segmentation results achieved by different methods on five COVID-19 cases. Red pixels represent normal lung areas, green pixels represent ground-glass opacities (GGO), and blue pixels represent lung consolidation. Some error-prone regions are denoted by yellow boxes.

addition, the GGO will transform to lung consolidation as the progression of the patient's condition, it is difficult for segmentation methods and physicians to distinguish lesions in this transformation process. However, the EMTN ensures that the whole lesion areas in the lung are correctly segmented, which can be reliably used for subsequent severity assessment of COVID-19.

5.2. Ablation study

In this section, we take the COVID-19 *vs.* NC classification target and multi-lesion segmentation as examples to evaluate the influence of MIL and MTL strategies on the performance of EMTN.

5.2.1. Influence of MIL strategy

To evaluate the influence of the explainable MIL strategy and its effectiveness, EM-Cls, the normal MIL method Gated-Attention, and the non-MIL method ResNet18-Voting are compared. Table 4 shows the comparison on evaluation metrics in terms of COVID-19 *vs.* NC classification target.

From Table 4, it can be learned that MIL methods (i.e., EM-Cls and Gated-Attention) yield better results in various metrics. One reason for these results is that non-MIL methods are susceptible to uneven distribution of lesions in CT images. This also proves that MIL methods can pay more attention to relationships between instances, which is helpful to improve the final expression abilities of instances when dealing with such weakly supervised problem. The proposed EM-Cls achieves the best classification performance in the MIL methods, with the ACC of 96.73%, the F1 Score of 95.89%, and the AUC of 96.98%, which is at least 1.39% higher than the metrics generated by Gated-Attention. These results reflect that the proposed explainable MIL strategy can further improve the classification performance and is more effective than normal MIL methods.

Furthermore, we analyze the performance of EMTN using different sizes of bags as inputs. Multi-lesion segmentation is a type of semantic segmentation task, where changing the bag size has little effect on the segmentation performance. Thus, Table 5 shows the classification performance corresponding to different bag sizes, and Table 7 shows the corresponding area measure $pAUC_c$. It can be observed that the performance of EMTN fluctuates with the changing bag sizes. The best classification performance is reached at the bag size of 100, and the $pAUC_c$ of the leftmost partial curve is higher than others. When the bag size is smaller than 100, the performance gradually improves as the bag size increases; and the performance tends to be saturated when larger than 80, the $pAUC_c$ of the leftmost partial curve under

Table 4
Evaluation of the explainable MIL strategy (bag size=100).

Method	ACC(%)	Precision(%)	Recall(%)	F1 Score(%)	AUC(%)
ResNet18-Voting	88.97 [85.52, 92.76]	85.85 [80.49, 92.03]	90.37 [85.14, 94.85]	88.05 [83.74, 92.02]	90.08 [86.30, 92.18]
Gated-Attention	94.48 [92.07, 97.24]	95.89 [92.28, 98.73]	93.24 [91.61, 97.08]	94.55 [83.74, 92.02]	94.63 [91.65, 97.08]
EM-Cls	96.73 [95.00, 98.80]	97.22 [95.09, 100.0]	94.59 [90.21, 98.14]	95.89 [93.06, 98.17]	96.98 [93.80, 98.99]

The upper and lower bounds of 95% confidence interval are shown in [·].

Table 5
Classification performance of EMTN under different bag sizes.

Bag size	ACC(%)	Precision(%)	Recall(%)	F1 Score(%)	AUC(%)
30	95.17 [93.10, 97.59]	93.99 [90.59, 98.08]	94.59 [90.44, 98.53]	94.29 [91.23, 97.01]	95.11 [91.20, 98.08]
50	96.55 [94.83, 98.62]	95.86 [92.95, 99.24]	95.80 [91.84, 99.20]	95.83 [93.63, 98.43]	96.78 [93.46, 98.79]
80	97.41 [96.02, 99.14]	96.05 [93.50, 100.0]	97.13 [93.76, 100.0]	96.59 [94.20, 98.59]	97.23 [94.50, 99.17]
100	98.62 [97.59, 100.0]	97.33 [95.18, 100.0]	98.65 [96.41, 100.0]	97.99 [95.86, 100.0]	98.90 [97.65, 100.0]
120	98.27 [97.23, 99.65]	98.63 [96.06, 100.0]	97.30 [94.36, 100.0]	97.96 [95.96, 99.59]	98.21 [96.14, 99.40]
150	97.93 [96.55, 99.66]	96.05 [92.60, 98.59]	98.65 [96.64, 100.0]	97.33 [94.94, 99.17]	97.72 [95.42, 99.84]

The upper and lower bounds of 95% confidence interval are shown in [·].

Table 6
Classification performance of EMTN using different instances selection strategies.

Bag size	Random selection		Front selection		Middle selection		Back selection	
	ACC(%)	AUC(%)	ACC(%)	AUC(%)	AUC(%)	AUC(%)	ACC(%)	AUC(%)
30	95.17 [93.10, 97.59]	95.11 [91.20, 98.08]	94.89 [92.12, 97.21]	94.99 [90.97, 98.15]	94.41 [91.59, 96.80]	94.61 [90.73, 97.86]	94.85 [92.28, 97.38]	94.94 [90.93, 97.90]
50	96.55 [94.83, 98.62]	96.78 [93.46, 98.79]	95.17 [92.53, 97.47]	95.37 [91.54, 98.57]	95.00 [92.46, 97.23]	94.94 [90.72, 98.32]	96.46 [94.01, 98.41]	96.54 [93.40, 98.98]
80	97.41 [96.02, 99.14]	97.23 [94.50, 99.17]	95.65 [93.14, 97.90]	95.87 [92.70, 98.09]	95.34 [92.83, 97.47]	95.60 [92.37, 98.21]	97.02 [94.90, 98.89]	97.15 [94.58, 98.99]
100	98.62 [97.59, 100.0]	98.90 [97.65, 100.0]	96.34 [94.01, 98.41]	96.44 [93.38, 98.65]	95.80 [93.58, 97.93]	95.89 [93.10, 98.31]	97.57 [95.36, 99.13]	97.81 [95.65, 99.41]
120	98.27 [97.23, 99.65]	98.21 [96.14, 99.40]	96.28 [93.80, 98.35]	96.42 [93.41, 98.69]	95.61 [93.21, 97.70]	95.79 [92.84, 98.24]	98.03 [96.29, 99.57]	98.06 [96.06, 99.47]
150	97.93 [96.55, 99.66]	97.72 [95.42, 99.84]	95.86 [93.58, 98.04]	96.16 [93.10, 98.42]	95.55 [93.24, 97.62]	95.70 [92.48, 98.22]	97.79 [95.82, 99.35]	97.94 [95.72, 99.33]

The upper and lower bounds of 95% confidence interval are shown in [·].

Table 7
Area measure $pAUC_c$ of EMTN under different bag sizes.

Bag size	$pAUC_c$(%)		
	i = 1	i = 2	i = 3
30	61.77	16.67	16.67
50	63.44	16.67	16.67
80	63.89	16.67	16.67
100	65.56	16.67	16.67
120	64.87	16.67	16.67
150	63.84	17.21	16.67

different bag sizes are higher than 63.84%. This indicates that the proposed EMTN needs at least 80 CT slices to obtain acceptable performance.

As mentioned in Section 4.1, a random selection strategy is used to select instances and construct bags. And we further analyze the influence of different selection strategies on the performance of EMTN. Table 6 shows the classification performance using a random selection strategy and three other selection strategies to create the bags. The three selection strategies include selecting a set of slices in the front of CT sequence, in the back of CT sequence, and in the middle of CT sequence. For the convenience of description, they are referred to as front selection, back selection and middle selection, respectively. From Table 6, it can be learned that the results obtained by using other three selection strategies are close to that obtained by using random selection strategy, and creating the bags of instances by random selection can make the performance optimal. The distribution of lesions in CT images of different patients is diverse and not uniform, and the bags are unfavorable for the representations of corresponding individuals when the instances are selected on certain parts of CT images. For example, in the case of small bag size, there may be extreme situation that no positive instances in the bags of positive individuals. These results illustrate that using a random selection strategy to create the bags is more reasonable than the other three selection strategies. Furthermore, the performance achieved by using the other three selection strategies is close to the optimal and acceptable, which indicates that the multi-instance assumption in this work is effective and can reduce the interference caused by different instances selection strategies.

Table 8
Evaluation of the weight-adaptive MTL strategy on classification task and segmentation task (bag size=100).

Method	Classification task					Segmentation task		
	ACC(%)	Precision(%)	Recall(%)	F1 Score(%)	AUC(%)	DC(%)	PPV(%)	SEN(%)
EM-Cls	96.73 [95.00, 98.80]	97.22 [95.09, 100.0]	94.59 [90.21, 98.14]	95.89 [93.06, 98.17]	96.98 [93.80, 98.99]	–	–	–
EM-Seg	–	–	–	–	–	94.29 ± 3.4	94.49 ± 5.0	94.08 ± 3.9
EMTN	98.62 [97.59, 100.0]	97.33 [95.18, 100.0]	98.65 [96.41, 100.0]	97.99 [95.86, 100.0]	98.90 [97.65, 100.0]	96.18 ± 3.7	96.26 ± 4.2	96.09 ± 3.8

Classification task results: the upper and lower bounds of 95% confidence interval are shown in [·].
Segmentation task results: the results are shown as mean ± standard deviation.

Table 9
Results for classification task and segmentation task with different λ (bag size=100).

$\lambda_{cls}/\lambda_{seg}$	Classification task					Segmentation task		
	ACC(%)	Precision(%)	Recall(%)	F1 Score(%)	AUC(%)	DC(%)	PPV(%)	SEN(%)
1/1	97.44 [95.72, 99.50]	96.16 [93.46, 99.58]	97.53 [94.30, 100.0]	96.84 [94.24, 98.95]	97.52 [94.92, 99.19]	95.13 ± 3.6	95.29 ± 4.5	94.97 ± 4.2
0.6/1	98.37 [97.34, 99.75]	98.61 [95.89, 100.0]	97.25 [96.03, 100.0]	97.93 [95.85, 99.67]	98.48 [96.67, 99.45]	95.94 ± 3.3	96.06 ± 4.4	95.82 ± 3.5
0.2/1	96.86 [95.13, 98.93]	95.95 [93.06, 99.36]	96.37 [92.78, 98.98]	96.16 [93.40, 98.37]	97.03 [93.90, 98.93]	95.42 ± 4.0	95.78 ± 4.7	95.07 ± 4.5
Learning	98.62 [97.59, 100.0]	97.33 [95.18, 100.0]	98.65 [96.41, 100.0]	97.99 [95.86, 100.0]	98.90 [97.65, 100.0]	96.18 ± 3.7	96.26 ± 4.2	96.09 ± 3.8

Classification task results: the upper and lower bounds of 95% confidence interval are shown in [·].
Segmentation task results: the results are shown as mean ± standard deviation.

5.2.2. Influence of MTL strategy

The proposed EMTN can simultaneously perform classification task and segmentation task in a weight-adaptive multi-task learning manner. We compare the performance of EMTN, EM-Cls, and EM-Seg to analyze the influence of MTL strategy. Table 8 summarizes the results of COVID-19 *vs*. NC classification and multi-lesion segmentation. As shown in Table 8, for classification task, MTL strategy can further improve the classification performance of EM-Cls. EMTN improves the ACC from 96.73% to 98.62%, the F1 Score from 95.89% to 97.99%, and the AUC from 96.98% to 98.90%. Meanwhile, for segmentation task, EMTN can achieve satisfactory segmentation results in terms of overall metrics with the influence of MTL strategy. In addition, Fig. 7 shows that EMTN generates more complete and detailed segmentation masks than EM-Seg. In general, the performance of EMTN based on the weight-adaptive MTL strategy is further improved on both classification of COVID-19 and multi-lesion segmentation tasks, which proves that the proposed MTL strategy can utilize task-related information to improve the performance on each task.

The multi-task loss function with adaptive weights is designed to actively adjust the contribution ratio of different tasks to the training of the network parameters. We further analyze the influence of trade-off factors λ in Eq. (3). As mentioned in Section 3.2, λ_{cls} and λ_{seg} denote the trade-off factors for classification task and segmentation task, respectively, which are learnable parameters modified during the iteration. Considering that the segmentation task is a fine-grained pixel-level task, and the classification task is a coarse-grained individual-level task, the segmentation task makes a greater contribution to the network parameter optimization during the training process. Specifically, we vary the ratio of λ_{cls} and λ_{seg} within $\{1/1, 0.6/1, 0.2/1\}$, and investigate the performance of EMTN when trade-off factors λ are different constants and learnable parameters. The comparison results are shown in Table 9, and area measure $pAUC_c$ for classification task with different λ are shown in Table 10.

From Tables 9 and 10, it can be observed that the proposed EMTN can further improve the performance on both tasks when λ are learnable parameters. Fig. 8 shows the learning curves of the trade-off factors. The initial values of λ_{cls} and λ_{seg} are both set as 1, and the values of λ are recorded after each iteration. λ_{cls} and λ_{seg} stabilize after about 40 iterations. The experiments prove that EMTN generates the best performance after around 49 iterations. The final values of λ_{cls} and λ_{seg} are approximately 0.5 and 0.84, respectively. Through the analysis of λ learning process, the contribution ratio of different tasks to the training of network parameters can be determined, which can avoid the tedious process of manually adjusting the trade-off factors. In addition, the larger factor assigned to the segmentation task also indicates that a fine-grained task provides more support for network parameter optimization than a coarse-grained task.

Table 10
Area measure $pAUC_c$ for classification task with different λ (bag size=100).

$\lambda_{cls}/\lambda_{seg}$	$pAUC_c$(%)		
	i = 1	i = 2	i = 3
1/1	64.18	16.67	16.67
0.6/1	65.14	16.67	16.67
0.2/1	63.69	16.67	16.67
Learning	65.56	16.67	16.67

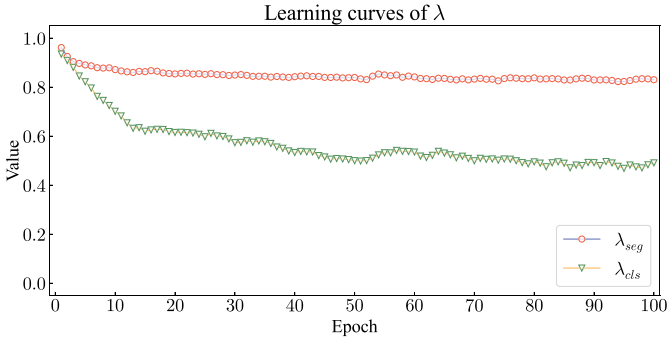

Fig. 8. Learning curves of trade-off factors λ. λ_{cls} and λ_{seg} respectively denote the trade-off factors for classification task and segmentation task.

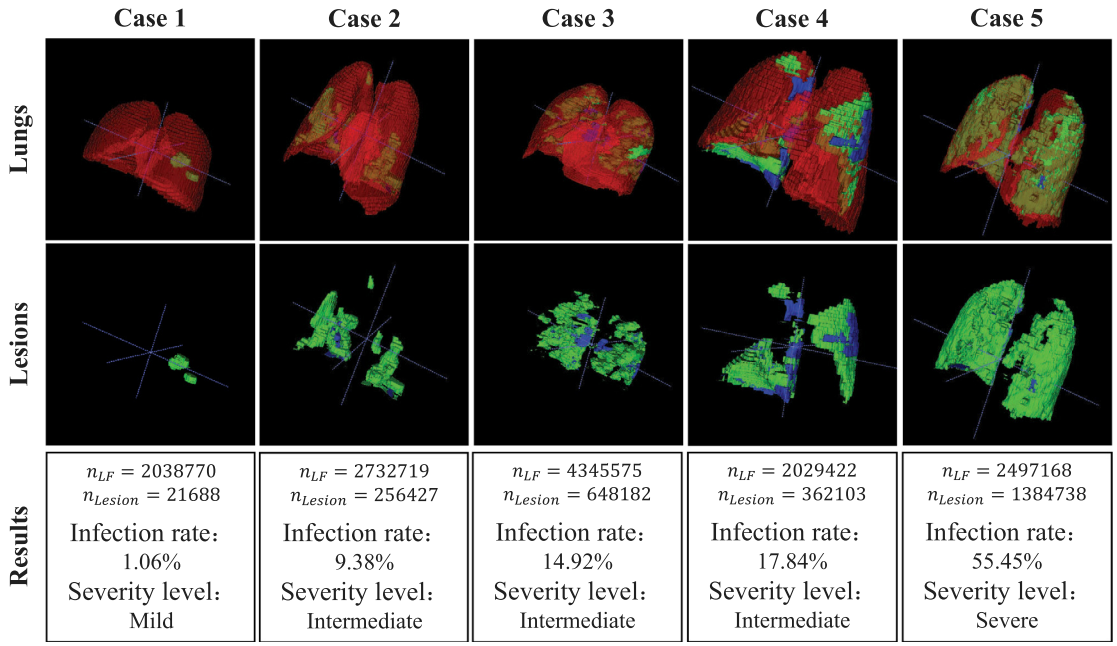

Fig. 9. Assessment results on five COVID-19 cases from mild, intermediate, and severe groups. For each case, we show the 3D visualization results of the whole lung and its lesion areas. The severity level definitions are as follows: less than three GGO lesions and lesion areas less than 5% of the entire lung fields is defined as mild; lesion areas more than 5% of the entire lung fields is defined as intermediate; lesion areas more than 40% of the entire lung fields is defined as severe.

5.2.3. Quantitative analysis of lesions and severity assessment of COVID-19

The quantitative analysis of lesions and severity assessment of COVID-19 are considered as the subsequent tasks of multi-lesion segmentation. Fig. 9 shows five assessment examples.

Specifically, the first, the second and the third rows of Fig. 9 present the segmentation results, the lesion areas and the assessment results, respectively. From the segmentation results and the separated lesion areas, it can be observed that the lesion areas of the entire lung fields from case 1 to case 5 are gradually increasing, while the infection rate of lung fields and the severity level of patients are unavailable. In the assessment results, the number of voxels in the lung fields and lesion areas, infection rate, and severity assessment of patients are shown. For example, the assessment results of case 3 suggest that the infection rate is 14.92% and the severity level of the patient is intermediate. Compared with direct observation, quantitative analysis and severity assessment can give the infection rate of lung fields and the severity level of patients, and these assessment results can serve as a reference for the diagnosis of COVID-19.

5.2.4. Explainability analysis of diagnosis results

As mentioned in Section 3.3, the proposed MIL strategy can deal with the weakly supervised problem and make EMTN have explainability by suggesting the key instances. We further investigate the explainability of our method by deriving the attention weights between the class token and each instance-level feature. The comparison of some key instances and non-key instances is shown in Fig. 10.

Specifically, we select the key instances based on the values of the attention weights, and the threshold is set to be 0.5. The instances with weights greater than the threshold are called key instances. Multiple experiments prove that these key instances usually account for about 20% or even less of the whole bag. As shown in Fig. 10, the first to third columns and the fourth to sixth columns are key instances and non-key instances, respectively; the second, the fourth and the last rows represent the segmentation masks of the first, the third and the fifth rows, respectively. For case 1, the weights of these six instances are

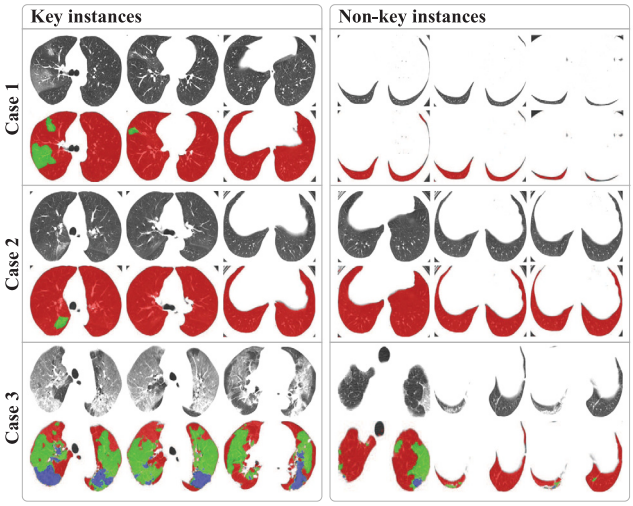

Fig. 10. Explainability analysis of diagnosis results of three cases by suggesting their key instances.

1.0, 0.5999, 0.5871, 0.0154, 0.0088, 0.0 from left to right; for case 2, the weights are 1.0, 0.6734, 0.6453, 0.0593, 0.0463, 0.0 from left to right; for case 3, the weights are 1.0, 0.8396, 0.7035, 0.1538, 0.0144, 0.0 from left to right. These weights are not normalized, and the larger weights mean that the corresponding instances have a greater influence on the diagnosis. It can be observed that the key instances of cases contain more lesion areas or complete lung fields than non-key instances, which can also roughly suggest the severity level of patients. Compared with key instances, the non-key instances contain many irrelevant areas, which have limited influence on the diagnosis of patients.

Furthermore, consistency is important for clinical diagnosis, and diagnostic decisions made by the same method are often influenced by similar images. To evaluate the consistency of the explainability part of the proposed framework, namely, to judge

Table 11
Results for classification task on COVID-19-CT & Radiopaedia dataset. (bag size=100).

Method	Training data	Testing data	ACC(%)	Precision(%)	Recall(%)	F1 Score(%)	AUC(%)
EM-Cls	CC-CCII	COVID-19-CT & Radiopaedia	88.37 [84.68, 92.34]	87.93 [82.57, 93.45]	87.29 [81.99, 92.86]	87.61 [82.69, 91.80]	89.17 [83.68, 93.18]
	COVID-19-CT & Radiopaedia	COVID-19-CT & Radiopaedia	94.35 [91.94, 97.18]	93.44 [89.83, 97.52]	94.99 [90.83, 98.38]	94.21 [90.98, 97.02]	95.07 [90.52, 98.21]
EMTN	CC-CCII	COVID-19-CT & Radiopaedia	89.92 [86.29, 93.95]	90.29 [85.48, 95.33]	89.05 [83.26, 94.14]	89.67 [85.60, 93.45]	90.12 [84.75, 94.56]

The upper and lower bounds of 95% confidence interval are shown in [·].

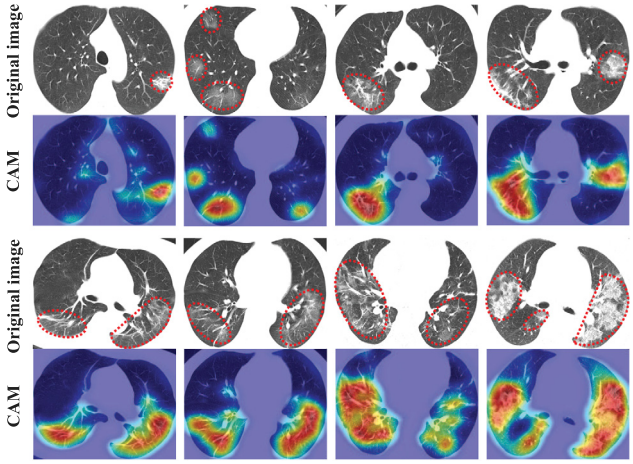

Fig. 11. Class activation maps of suggested key instances obtained by Grad-CAM++.

whether the suggested key instances can provide the posterior test of the diagnosis results, we give more explainable analysis from the aspect of semantic comprehension. We use Grad-CAM++ [40] to obtain class activation maps (CAMs) of several key instances from different patients. As shown in Fig. 11, the first and third rows are key instances (the lesion regions have been marked with dashed lines), the second and fourth rows are the corresponding class activation maps. It can be observed that the class activation maps of suggested key instances show the approximate locations of lesions, which indicates the key instances suggested by the explainability part of this framework are roughly consistent with the explainable analysis from the semantic perspective. These results imply that the proposed explainable MIL can provide the posterior test of the diagnosis results, and indicate the concern level of the network for different instances.

5.3. Additional study

To further evaluate the proposed method, we analyze the performance of EMTN and EM-Cls on other open-source datasets referred to as COVID-19-CT & Radiopaedia (i.e., COVID-19-CT dataset and common pneumonia patients collected from Radiopaedia). Tables 11 and 12 show the classification results and area measure $pAUC_c$ of EMTN and EM-Cls on the COVID-19-CT & Radiopaedia dataset, respectively. Fig. 12 shows three severity assessment examples based on multi-lesion segmentation masks generated by EMTN. It should be noted that due to the lack of ground truth for segmentation in the COVID-19-CT & Radiopaedia data, the dataset has not participated in the training stage of EMTN. The partial results in Table 11, Table 12, and Fig. 12 are obtained by directly using the EMTN which trained on the CC-CCII dataset for testing this dataset. For clarity, we illustrate the training and test data in the tables.

Table 12
Area measure $pAUC_c$ for classification task on COVID-19-CT & Radiopaedia dataset (bag size=100).

Method	Training data	Testing data	$pAUC_c$(%)		
			i = 1	i = 2	i = 3
EM-Cls	CC-CCII	COVID-19-CT & Radiopaedia	54.06	18.92	16.19
	COVID-19-CT & Radiopaedia	COVID-19-CT & Radiopaedia	61.03	17.37	16.67
EMTN	CC-CCII	COVID-19-CT & Radiopaedia	56.15	17.48	16.49

As shown in Table 11, without using COVID-19-CT & Radiopaedia dataset for training, the classification accuracy achieved by EMTN and EM-Cls on this dataset are 89.92% and 88.37%, respectively, and AUCs are about 90%. After training EM-Cls with COVID-19-CT & Radiopaedia dataset, the evaluation metrics are on average 7.20% higher than that of EM-Cls without using COVID-19-CT & Radiopaedia dataset for training. The ACC (94.35%) and AUC (95.07%) have 6.77% and 6.62% improvement, respectively. From Table 12, it can be observed that the ROC curve of EM-Cls trained with the COVID-19-CT & Radiopaedia dataset goes up faster while staying left, and it has a higher value of TPR than that trained with the CC-CCII dataset. In Fig. 12, the first, the second, and the third rows present the segmentation results, the lesion areas and the assessment results, respectively. The assessment results of these cases suggest the infection rate and the severity level of the patients. From case 1 to case 3, their assessment results are mild, intermediate, intermediate, and their infection rates are 0.91%, 6.12%, 9.39%.

Considering the influence of the differences between datasets on the model, though the classification results of EMTN and EM-Cls decline without using the COVID-19-CT & Radiopaedia as training data, these results are acceptable. In the situation of using the COVID-19-CT & Radiopaedia as training data, EM-Cls as a variant of EMTN can achieve satisfactory results. For the segmentation branch in EMTN, it accesses a large number of pixels and learns to correctly predict each pixel to the corresponding semantic category during the training process. Therefore, generating segmentation masks for COVID-19-CT & Radiopaedia dataset by EMTN trained on the CC-CCII dataset is less affected by the differences between datasets, and the generated masks are also relatively accurate. The above analysis indicates that the proposed method is applicable to the new datasets and can achieve satisfactory results.

5.4. Discussion on the limitation

This work is oriented towards an actual problem, that is, the auxiliary diagnosis of COVID-19. The effectiveness of the proposed framework has been verified by extensive experiments based on real-world datasets, yet there are still some limitations in our work. In this section, we discuss the limitations of this work and the gap with practical clinical applications.

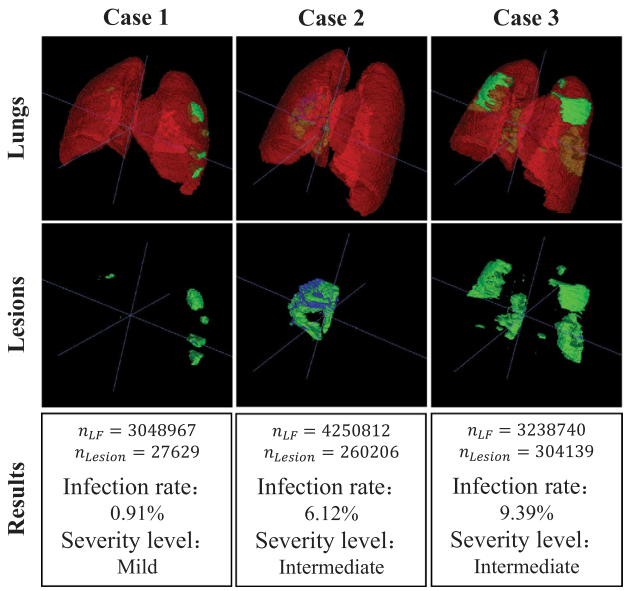

Fig. 12. Assessment results on three cases in COVID-19-CT & Radiopaedia dataset.

As far as clinical data is concerned, data quality, data standards, and data amounts are all issues that need to be considered in the transition from algorithm to practical application. The data preprocessing procedures and multi-instance learning strategy adopted in this work can alleviate the problem of uneven data quality and diverse data standards, which is achieved by obtaining regions of interest and randomly selecting instances to construct a bag of instances and represent an individual. The real-world datasets of CT images used in this work have certain limitations, part of CT images have no ground truth for segmentation task, and the severity of COVID-19 has to be assessed through a robust voxel-wise analysis method. The proposed framework will be further perfected if more well-labeled datasets are available, such as those used for the severity assessment of COVID-19. In addition, considering the complexity of COVID-19, for some special cases, such as asymptomatic infections without obvious CT imaging features, decision-making cannot be performed only based on CT images. In this situation, the diagnosis of COVID-19 should be made under more comprehensive tests which may include RT-PCR or other clinical examinations. It also inspired that the information generated by these clinical examinations can be adopted as auxiliary diagnosis indexes to help construct a more robust computer-aided diagnosis system.

On the other hand, the interpretability of auxiliary diagnosis methods based on deep learning is weak, while the process of clinical diagnosis requires rigorous evidence. The black-box nature is one of the main reasons which limits the wide application of fully automated medical artificial intelligence (AI). In the European "Artificial Intelligence Act", some clear guidance on the use of medical AI has been already provided, which constitutes a binding legal framework for the use of medical AI. This is the protection of human rights in the context of medical AI development. Stoeger et al. [41] point out that human oversight and explainability are required in medical AI, namely, one AI system must be explainable to be used in medicine. It not only coincides with the requirements of European fundamental rights, but also with the demands of computer science. Therefore, the importance of explainable medical AI is self-evident. Though the proposed explainable multi-instance learning can give the explainability analysis of diagnosis results by suggesting the key instances, it is not explainable in the mathematical sense but in the clinical sense. Furthermore, it also requires human oversight and a more complete explanation. Interactive machine learning with the "human in the loop" could be a potential solution to this limitation of AI [42,43]. It should be noted that physicians/radiologists have conceptual understanding and experience that no AI can fully learn. Combined with the conceptual understanding and experience of physicians/radiologists, the interactive machine learning with the "human in the loop" can find the underlying explanatory factors for AI, which ensure that decisions made by AI can be human-controlled and clinically justified. This human-in-the-loop machine learning will be explored in our future work.

6. Conclusion

In this paper, we construct an integrated framework for segmenting lesion areas and diagnosing COVID-19 from CT images. It takes the explainable multi-instance multi-task network (EMTN) as the core, and the lesion quantification and severity assessment as important components. The EMTN can make proper use of task-related information to further improve the performance on diagnosis and segmentation, and it also has EM-Seg (EMTN with only the segmentation branch) and EM-Cls (EMTN with only the classification branch) two variants. These two variants can be respectively employed to perform diagnosis and segmentation, which improve the flexibility of EMTN. An explainable multi-instance learning strategy is proposed in EMTN for explainability analysis of diagnosis results, and a weight-adaptive multi-task learning strategy is proposed for the coordination between both tasks. Considering that the evaluation of patient status needs more detailed information, the lesion quantification and severity assessment are adopted as the subsequent tasks of multi-lesion segmentation. The experimental results show that the proposed EMTN has better performance over several mainstream methods on the diagnosis and multi-lesion segmentation of COVID-19, and can provide explicable diagnosis results that enhance the explainability of the network. Moreover, the proposed framework gives the assessment of COVID-19 patient status as an extra reference for the auxiliary diagnosis of COVID-19, including the lung infection rate and the severity level of patients.

CRediT authorship contribution statement

Minglei Li: Conceptualization, Methodology, Validation, Writing – original draft. **Xiang Li:** Conceptualization, Methodology, Writing – review & editing. **Yuchen Jiang:** Conceptualization, Methodology, Writing – review & editing. **Jiusi Zhang:** Conceptualization, Methodology. **Hao Luo:** Supervision, Writing – review & editing, Funding acquisition. **Shen Yin:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] T.G. Ksiazek, D. Erdman, C.S. Goldsmith, S.R. Zaki, T. Peret, S. Emery, S. Tong, C. Urbani, J.A. Comer, W. Lim, et al., A novel coronavirus associated with severe acute respiratory syndrome, N. Engl. J. Med. 348 (20) (2003) 1953–1966.

[2] R.J. De Groot, S.C. Baker, R.S. Baric, C.S. Brown, C. Drosten, L. Enjuanes, R.A. Fouchier, M. Galiano, A.E. Gorbalenya, Z.A. Memish, et al., Commentary: Middle east respiratory syndrome coronavirus (mers-cov): announcement of the coronavirus study group, J. Virology 87 (14) (2013) 7790–7792.

[3] C. Wang, P.W. Horby, F.G. Hayden, G.F. Gao, A novel coronavirus outbreak of global health concern, Lancet 395 (10223) (2020) 470–473.
[4] X. Xie, Z. Zhong, W. Zhao, C. Zheng, F. Wang, J. Liu, Chest CT for typical coronavirus disease 2019 (COVID-19) pneumonia: Relationship to negative RT-PCR testing, Radiology 296 (2) (2020) E41–E45.
[5] U. Muhammad, M.Z. Hoque, M. Oussalah, A. Keskinarkaus, T. Seppänen, P. Sarder, SAM: Self-augmentation mechanism for COVID-19 detection using chest X-ray images, Knowl.-Based Syst. (2022) 108207.
[6] C. Li, Y. Yang, H. Liang, B. Wu, Transfer learning for establishment of recognition of COVID-19 on CT imaging using small-sized training datasets, Knowl.-Based Syst. 218 (2021) 106849.
[7] H. Shi, X. Han, N. Jiang, Y. Cao, O. Alwalid, J. Gu, Y. Fan, C. Zheng, Radiological findings from 81 patients with COVID-19 pneumonia in wuhan, China: A descriptive study, Lancet Infect. Dis. 20 (4) (2020) 425–434.
[8] A.A. Ardakani, A.R. Kanafi, U.R. Acharya, N. Khadem, A. Mohammadi, Application of deep learning technique to manage COVID-19 in routine clinical practice using CT images: Results of 10 convolutional neural networks, Comput. Biol. Med. 121 (2020) 103795.
[9] H.X. Bai, R. Wang, Z. Xiong, B. Hsieh, K. Chang, K. Halsey, T.M.L. Tran, J.W. Choi, D.-C. Wang, L.-B. Shi, et al., Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT, Radiology 296 (3) (2020) E156–E165.
[10] Z. Han, B. Wei, Y. Hong, T. Li, J. Cong, X. Zhu, H. Wei, W. Zhang, Accurate screening of COVID-19 using attention-based deep 3D multiple instance learning, IEEE Trans. Med. Imaging 39 (8) (2020) 2584–2594.
[11] S. Chaganti, P. Grenier, A. Balachandran, G. Chabin, S. Cohen, T. Flohr, B. Georgescu, S. Grbic, S. Liu, F. Mellot, et al., Automated quantification of CT patterns associated with COVID-19 from chest CT, Radiology: Artif. Intell. 2 (4) (2020) e200048.
[12] W. Xie, C. Jacobs, J.-P. Charbonnier, B. Van Ginneken, Relational modeling for robust and efficient pulmonary lobe segmentation in CT scans, IEEE Trans. Med. Imaging 39 (8) (2020) 2664–2675.
[13] K. Zhang, X. Liu, J. Shen, Z. Li, Y. Sang, X. Wu, Y. Zha, W. Liang, C. Wang, K. Wang, et al., Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography, Cell 181 (6) (2020) 1423–1433.
[14] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, L. Van Gool, Multi-task learning for dense prediction tasks: A survey, IEEE Trans. Pattern Anal. Mach. Intell. (2021).
[15] M. Abdel-Basset, V. Chang, H. Hawash, R.K. Chakrabortty, M. Ryan, FSS-2019-nCov: A deep learning architecture for semi-supervised few-shot segmentation of COVID-19 infection, Knowl.-Based Syst. 212 (2021) 106647.
[16] Y.-H. Wu, S.-H. Gao, J. Mei, J. Xu, D.-P. Fan, R.-G. Zhang, M.-M. Cheng, Jcs: An explainable COVID-19 diagnosis system by joint classification and segmentation, IEEE Trans. Image Process. 30 (2021) 3113–3126.
[17] G. Chassagnon, M. Vakalopoulou, E. Battistella, S. Christodoulidis, T.-N. Hoang-Thi, S. Dangeard, E. Deutsch, F. Andre, E. Guillo, N. Halm, et al., AI-driven quantification, staging and outcome prediction of COVID-19 pneumonia, Med. Image Anal. 67 (2021) 101860.
[18] M. Vakalopoulou, G. Chassagnon, N. Bus, R. Marini, E.I. Zacharaki, M.-P. Revel, N. Paragios, AtlasNet: Multi-atlas non-linear deep networks for medical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 658–666.
[19] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[20] D.-P. Fan, T. Zhou, G.-P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, L. Shao, Inf-net: Automatic COVID-19 lung infection segmentation from ct images, IEEE Trans. Med. Imaging 39 (8) (2020) 2626–2637.
[21] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, et al., Chexnet: Radiologist-level pneumonia detection on chest X-rays with deep learning, 2017, arXiv preprint arXiv:1711.05225.
[22] J. Wang, Y. Bao, Y. Wen, H. Lu, H. Luo, Y. Xiang, X. Li, C. Liu, D. Qian, Prior-attention residual learning for more discriminative COVID-19 screening in CT images, IEEE Trans. Med. Imaging 39 (8) (2020) 2572–2583.
[23] O. Çiçek, A. Abdulkadir, S.S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: learning dense volumetric segmentation from sparse annotation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2016, pp. 424–432.
[24] X. Wang, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, W. Liu, C. Zheng, A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT, IEEE Trans. Med. Imaging 39 (8) (2020) 2615–2625.
[25] W.M. Shaban, A.H. Rabie, A.I. Saleh, M. Abo-Elsoud, A new COVID-19 patients detection strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier, Knowl.-Based Syst. 205 (2020) 106270.
[26] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.
[27] S. Wang, Y. Zha, W. Li, Q. Wu, X. Li, M. Niu, M. Wang, X. Qiu, H. Li, H. Yu, et al., A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis, Eur. Respir. J. 56 (2) (2020).
[28] Y. Song, S. Zheng, L. Li, X. Zhang, X. Zhang, Z. Huang, J. Chen, R. Wang, H. Zhao, Y. Zha, et al., Deep learning enables accurate diagnosis of novel coronavirus (COVID-19) with CT images, IEEE/ACM Trans. Comput. Biol. Bioinform. (2021).
[29] G. Bradski, A. Kaehler, Learning OpenCV: Computer Vision with the OpenCV Library, " O'Reilly Media, Inc.", 2008.
[30] S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, M. Cai, J. Yang, Y. Li, X. Meng, et al., A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19), Eur. Radiol. (2021) 1–9.
[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[32] X. Li, Y. Jiang, M. Li, S. Yin, Lightweight attention convolutional neural network for retinal vessel image segmentation, IEEE Trans. Ind. Inf. 17 (3) (2020) 1958–1967.
[33] L. Chen, X. Jiang, X. Liu, Z. Zhou, Logarithmic norm regularized low-rank factorization for matrix and tensor completion, IEEE Trans. Image Process. 30 (2021) 3434–3449.
[34] D. Alexey, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, An image is worth 16×16 words: Transformers for image recognition at scale, in: Proceedings of the International Conference on Learning Representations, Lisbon, Portuagal, 2021, pp. 7–8.
[35] S. Shakouri, M.A. Bakhshali, P. Layegh, B. Kiani, F. Masoumi, S. Ataei Nakhaei, S.M. Mostafavi, COVID19-CT-dataset: an open-access chest CT image repository of 1000+ patients with confirmed COVID-19 diagnosis, BMC Res. Notes 14 (1) (2021) 1–3.
[36] A.M. Carrington, P.W. Fieguth, H. Qazi, A. Holzinger, H.H. Chen, F. Mayr, D.G. Manuel, A new concordant partial AUC and partial C statistic for imbalanced data in the evaluation of machine learning algorithms, BMC Med. Inform. Decis. Mak. 20 (1) (2020) 1–12.
[37] T.G. Dietterich, Ensemble methods in machine learning, in: International Workshop on Multiple Classifier Systems, Springer, 2000, pp. 1–15.
[38] M. Ilse, J. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: International Conference on Machine Learning, PMLR, 2018, pp. 2127–2136.
[39] Z. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, J. Liang, Unet++: A nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2018, pp. 3–11.
[40] A. Chattopadhay, A. Sarkar, P. Howlader, V.N. Balasubramanian, Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, in: 2018 IEEE Winter Conference on Applications of Computer Vision, WACV, IEEE, 2018, pp. 839–847.
[41] K. Stöger, D. Schneeberger, A. Holzinger, Medical artificial intelligence: The European legal perspective, Commun. ACM 64 (11) (2021) 34–36.
[42] E. Sorantin, M.G. Grasser, A. Hemmelmayr, S. Tschauner, F. Hrzic, V. Weiss, J. Lacekova, A. Holzinger, The augmented radiologist: Artificial intelligence in the practice of radiology, Pediatr. Radiol. (2021) 1–13.
[43] Y. Jiang, X. Li, H. Luo, S. Yin, O. Kaynak, Quo vadis artificial intelligence? Discov. Artif. Intell. 2 (1) (2022) 1–19.