Biomedical Signal Processing and Control 91 (2024) 105937

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

TransDD: A transformer-based dual-path decoder for improving the performance of thoracic diseases classification using chest X-ray

Xiaoben Jiang [a], Yu Zhu [a,*], Yatong Liu [a], Gan Cai [a], Hao Fang [b,c,*]

[a] School of Information Science and Technology, East China University of Science and Technology, Shanghai 200237, People's Republic of China
[b] Department of Anesthesiology, Shanghai Geriatric Medical Center, Shanghai 201100, People's Republic of China
[c] Department of Anesthesiology, Zhongshan Hospital, Fudan University, Shanghai 200032, People's Republic of China

ARTICLE INFO

Keywords:
Chest X-ray image
Dual-path decoder
Thoracic diseases classification
Classification attention block

ABSTRACT

Manually and accurately detecting thoracic diseases from CXR images is a time-consuming task that requires experienced radiologists. Therefore, automated thoracic diseases classification has great significance. However, most existing methods solely leverage the feature maps extracted from CXR images to classify thoracic diseases, without effectively connecting the correlation between the local discriminative lesion features and their corresponding labels. To address this issue, we innovatively introduce a learnable label embedding as queries to detect and match class-related features from the feature maps, and then processed by a novel Transformer-based dual-path decoder (TransDD) to facilitate interaction. The proposed TransDD is comprised of three key components: spatial reduction attention (SRA), dual-path attention (DPA), and feature enhancement module (FEM). SRA is employed in simplifying the complexity of self-attention, while DPA is specifically designed to connect the explicit correlation between the features and labels. Moreover, FEM is used to boost the expressiveness of local features. Subsequently, the classification attention block is utilized to balance two classification scores based on the feature output and label output, respectively. The proposed TransDD-PVT attained SOTA performance on the ChestX-ray14 dataset, achieving a mean area under the receiver operating characteristic (AUC) of 83.1% across all 14 classes. Also, our method achieves 94.31% accuracy and 93.31% sensitivity on three-class classifications. Extensive experiments conducted on several datasets demonstrate the powerful ability of our TransDD to improve the performance of thoracic diseases classification. It can serve as a plug-and-play structure to improve the classification performance of both CNNs and recent Transformer-based backbones.

1. Introduction

Thoracic diseases are serious health problems in the lives of people [1]. The Chest X-ray (CXR) is a diagnostic examination that is painless and non-invasive, and it has gained widespread usage in screening various thoracic diseases [2,3]. During the devastating COVID-19 epidemic that has caused serious health and economic consequences, CXR has played a crucial role in assisting clinical diagnosis [4,5]. However, CXR images are almost analyzed through radiologists' visual inspection which requires a high degree of skills and concentration. In contrast, many countries lack experienced radiologists who can read CXR images accurately [6]. Hence, an automated computer-aided diagnosis (CAD) of thoracic diseases is of great significance. Recognizing its significance, Wang et al. [7] first constructed the ChestX-ray14 dataset to evaluate the automated algorithms for CAD of thoracic diseases. After that, [8] designed a deep-learning pipeline for the diagnosis of pneumonia, and also constructed a CXR dataset (CC-CXRI) which is the largest multi-clinical scene CXR images dataset around the world.

With the continuous development of deep convolutional neural networks (DCNN) [9–11], researchers can mine available information from large-scale medical data. Benefiting from two large CXR datasets [7,8], various CNN-based methods [1,3,11–16] were employed for the diagnosis of thoracic diseases. However, most prior works remain some faultinesses. The main challenges in the field of thoracic diseases classification are common as follows: (1) As shown in Fig. 1(a)-(c), the lesion regions of thoracic diseases could have obvious scale variance and different locations of the lung field. The region of effusion in Fig. 1(a) is much smaller than in Fig. 1(c). (2) Multi-label diseases may appear in a

* Corresponding authors at: School of Information Science and Technology, East China University of Science and Technology, Shanghai 200237, People's Republic of China (Y. Zhu).
E-mail addresses: zhuyu@ecust.edu.cn (Y. Zhu), drfanghao@163.com (H. Fang).

https://doi.org/10.1016/j.bspc.2023.105937
Received 14 April 2023; Received in revised form 12 October 2023; Accepted 29 December 2023
Available online 13 January 2024
1746-8094/

single CXR image. For instance, there are Effusion, Infiltrate, and Mass in Fig. 1(b). (3) The existing research works lack the ability to comprehensively capture global lesion information. Moreover, most works solely leverage the feature maps extracted from CXR images to classify thoracic diseases, without effectively connecting the correlation between the local discriminative lesion features and corresponding labels.

Recently, the success of Transformer [17] in image classification [18–20] has been impressive. The ability to capture the information of the global image through the self-attention mechanism is one of the keys to success [21]. In addition, another advantage of the Transformer is its cross-attention mechanism, which can establish cross-modal connections. Motivated by the nomenclature used in Transformer, we introduce a learnable label embedding as queries to detect and match class-related features from the feature maps that are set as key and value. A novel Transformer-based dual-path decoder (TransDD) framework that can address the above challenges is comprised of the feature decoder and label decoder. In detail, there are three key components: spatial reduction attention (SRA), dual-path attention (DPA), and feature enhancement module (FEM). By simulating self-attention, our SRA can capture variances in appearance, location, and scale of the lesion regions in CXR images. Meanwhile, our SRA can reduce the complexity of the global self-attention. Moreover, the DPA can establish the connection between local discriminative features and the corresponding label. In addition, a FEM is employed to boost the expressiveness of local features. After that, we also designed a classification attention block to balance two classification scores based on feature output and label output, respectively. Note that our TransDD can serve as a plug-and-play structure to improve the performance of both CNNs (ie., EffNet [22], ResNet [14], and DenseNet [15]) and recent Transformer-based backbones (ie., ViT [18], PVT [19], and Swin Transformer [20]). To evaluate the effectiveness of our TransDD, extensive experiments were conducted on two tasks, multi-label and multi-class thoracic diseases classification. Compared with the state-of-the-art baselines, sufficient results demonstrated that the proposed TransDD has the remarkable capability to promote thoracic diseases classification. The main contributions of this paper are summarized as follows:

1. To connect the correlation between the local discriminative lesion features and the corresponding labels of thoracic diseases, we innovatively introduce a learnable label embedding as queries to detect and match class-related features from the feature maps.

2. We propose a novel Transformer-based dual-path decoder (TransDD), which is comprised of the feature decoder and label decoder. In detail, there are three key components: spatial reduction attention, dual-path attention, and feature enhancement module. Our TransDD can serve as a plug-and-play structure to enhance the classification performance of both CNNs and recent Transformer-based backbones.

3. After that, classification attention is designed to balance two classification scores based on feature output and label output

4. We verify the effectiveness of our TransDD with comprehensive experiments on the multi-label and multi-class classification of thoracic diseases. Sufficient experiments demonstrate that our proposed TransDD framework can bring a significant boost on the comparative backbone for thoracic diseases classification.

2. Related work

2.1. Cnn-based computer-aided diagnosis of thoracic diseases

With the continuous development and progress of medical image processing, more and more medical images need to be interpreted by doctors, which has gradually become a hot topic and challenge [23]. Doctors may have interpretation errors due to inexperience or fatigue, which are prone to false positive and false negative results [24]. In this situation, the emergence of computer-aided diagnosis (CAD) has finally become the demand of the times, which can significantly enhance the precision of diagnosis and offer efficient decision-making aid to physicians. CXR is an overlapping image of the human structure, which makes it easy to cover up local lesions and causes misdiagnosis. Hence, CAD in CXR images is widely employed for thoracic diseases [25]. To improve the performance of thoracic disease classification, many researchers employed various algorithms based on DCNNs. Wang et al. [7] compared several classic CNN architectures (i.e., AlexNet [11], VGGNet [12], GoogLeNet [13], and ResNet [14]) for the diagnosis of 14 thoracic diseases. [8] and [26] fine-tuned a DenseNet-121 [15] model for the CXR image classification on the CC-CXRI dataset and ChestX-ray14 dataset, respectively. Inspired by the attention mechanism that has been widely utilized in the realm of computer vision (CV), Guan et al. [16] proposed a highly adaptive category-wise residual attention module that can easily be incorporated into any feature embedding network, allowing for seamless end-to-end multi-label CXR image classification training. [1] designed a triple-attention learning (A^3 Net) model, which contains channel-wise, element-wise, and scale-wise attention learning for CAD of thoracic diseases. Thorax-net [27] consists of an attention branch that exploits the correlation between the locations of pathological abnormalities and class labels via analyzing the feature maps. In addition, Chen et al. [3] introduced a groundbreaking framework called Semantic Similarity Graph Embedding (SSGE), which meticulously investigates the semantic similarities existing within images to enhance the visual feature embedding. [28] presents an optimized ensemble framework for solving multi-label classification on long-tailed chest X-ray data. Jin et al. [29] proposed a new dual-weighted metric loss function for multi-label chest X-ray images. However, these methods could not capture variances in appearance, location, and scale of the lesion regions in CXR images, and failed to consider the connection

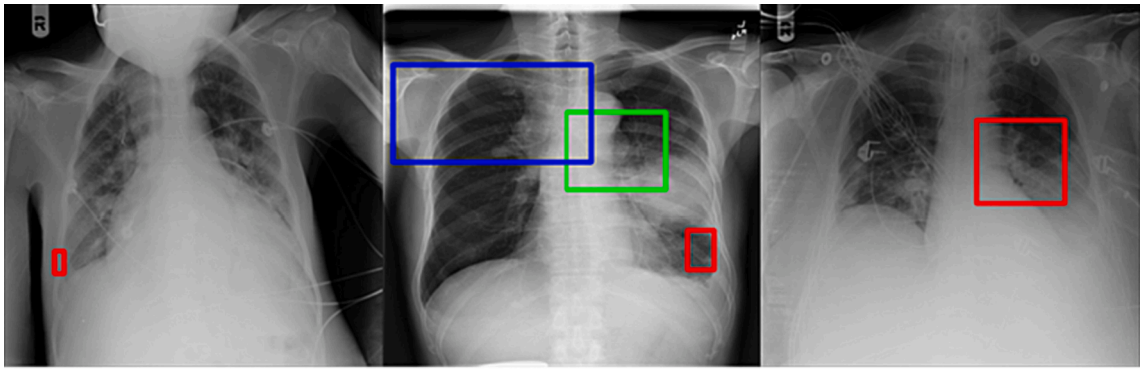

Fig. 1. Three CXR images from ChestX-ray14 with lesion regions labeled on the ground truth. Red, green, and blue bounding boxes represent effusion, infiltrate, and mass, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

between local discriminative diseased features and the corresponding label of thoracic diseases.

2.2. Transformer in image classification

Transformer [17] was first proposed for natural language processing (NLP) tasks [30–32] which had achieved state-of-the-art performance. Inspired by the achievements of the Transformer in the field of natural language processing (NLP), several researchers [18–21] pursued the integration of the Transformer into the realm of computer vision (CV). Vision Transformers (ViT) was first proposed in [18]. An image was split into non-overlapping static image patches as tokens and then fed into a stacked Transformer architecture for classification. However, the drawback of ViT is its reliance on extensive training datasets like JFT-300 M, containing 300 M images, despite its remarkable accuracy. Hence, Wang et al. [19] presented a Pyramid Vision Transformer (PVT) that can reduce the resolution of feature maps by incorporating the pyramid structure from CNN. Liu et al. [20] proposed a new vision Swin Transformer which produces a hierarchical feature representation with a linear computational complexity. Yang et al. [21] proposed a focal Transformer that applied either coarse-grained global attention or fine-grained local attention to reduce the memory cost. In addition, Jamali et al. [33] proposed a local window attention transformer and Zhao et al. [34] designed a multi-attention Transformer for image classification. The self-attention and decoder mechanism are the keys to success. Inspired by that, we propose a novel Transformer-based dual-path decoder (TransDD) framework for thoracic diseases classification.

3. Materials and method

3.1. 3.1 Datasets

Multi-label and multi-class classification are two common tasks in thoracic diseases classification. In multi-label classification, each input can have multi-output classes, while each input will have only one output class in multi-class classification. We extensively validate our proposed TransDD framework on both multi-label and multi-class classification of thoracic diseases.

Multi-label classification of thoracic diseases. The ChestX-ray14 dataset (ChestX-ray14) [7] published by the NIH, is widely regarded as the most commonly utilized benchmark in the field of automatic multi-label CXR image analysis. It consists of 112,120 frontal-view CXR images obtained from 30,805 individual patients, and each image is labeled for up to 14 diseases or "No Finding". In addition, China Consortium of Chest X-ray Image Investigation also constructed large CXR datasets. First, Sun Yat-sen University (SYSU) dataset [8] (including 120,012 CXR images) consists of patients from hospital visits. SYSU-PE [8] (including 42,402 CXR images) is another dataset containing additional patients who underwent a routine annual physician examination for external validation.

Multi-class classification of thoracic diseases. Wang et al. [8] also released another dataset (CC-CXRI-P), consisting of 7,921 CXR images for detecting viral pneumonia (including COVID-19 pneumonia), other types of pneumonia, and normal controls.

On the whole, we employed the ChestX-ray14, SYSU, and SYSU-PE datasets for multi-label classification of thoracic diseases, while utilizing the CC-CXRI-P dataset for multi-class classification.

3.2. Framework

The overall framework for thoracic diseases classification is depicted in Fig. 2. Given an input CXR image X with a set of categories of thoracic diseases, our framework can predict whether each pathology is present. Before performing SRA and DPA, we need to reshape the feature maps $X \in \mathbb{R}^{h\times w\times d}$ extracted by the backbone into a sequence of flattened 2D patches $F \in \mathbb{R}^{hw\times d}$. Here, h, w, and d represent the length, width, and dimension of the feature map. And, hw means the number of patches is $(h \times w)$. The backbone could be recently developed Vision Transformer backbones (ie., ViT [18], PVT [19], Swin Transformer [20]) or classical CNN backbones (ie., ResNet [14], DenseNet [15]). Meanwhile, we innovatively introduce a randomly initialized two-dimensional matrix $L \in \mathbb{R}^{N\times d}$ as a learnable label embedding, where N is the number of categories. Then we sent the label embedding L to our TransDD to detect and match class-related features from the feature F. The proposed TransDD can efficiently establish the connection between local discriminative diseased features and the corresponding label of thoracic diseases. Finally, we perform the classification attention block to

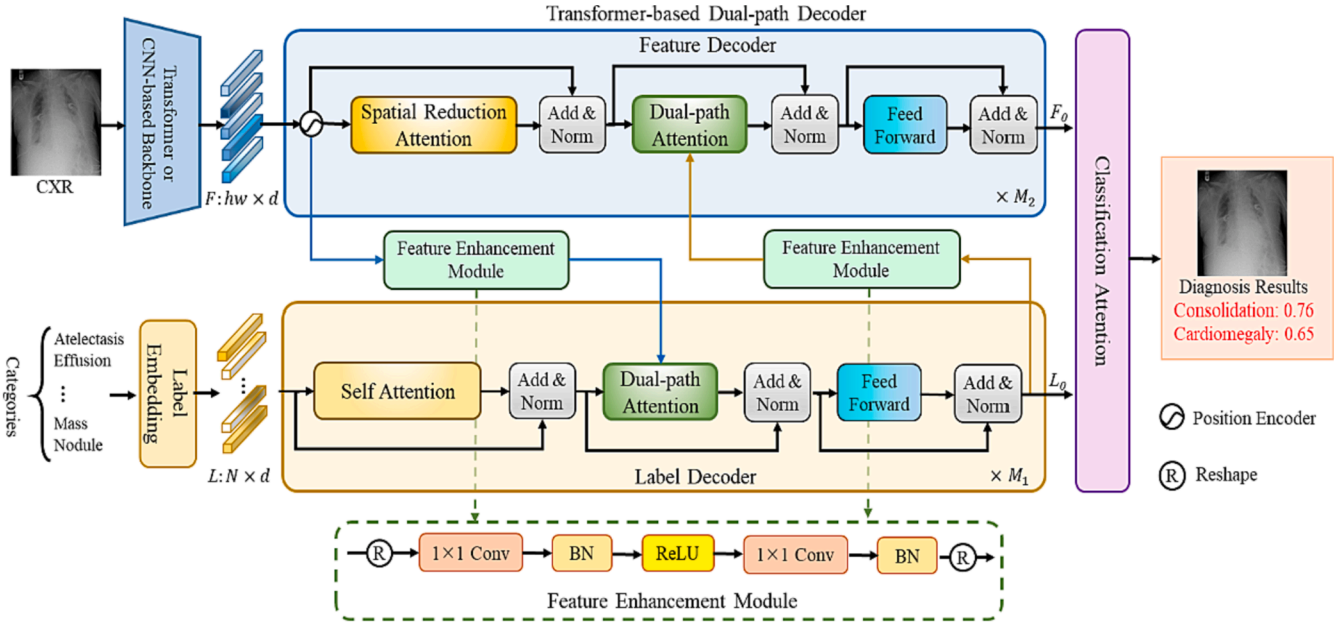

Fig. 2. The overall framework for thoracic diseases classification. After extracting spatial features from the backbone, a learnable label embedding is sent to our Transformer-based dual-path decoder (TransDD) to detect and match class-related features from the feature maps. The classification attention is then used to balance two classification scores based on feature output and label output.

balance the two classification scores yielded by feature output and label output, respectively.

3.2.1. Transformer-based dual-path decoder block

After yielding the features $F \in \mathbb{R}^{hw \times d}$ and a learnable label embedding $L \in \mathbb{R}^{N \times d}$, we design a Transformer-based dual-path decoder (TransDD) which consists of a feature decoder and a label decoder, to efficiently establish the connection between local discriminative diseased features and the corresponding label of thoracic diseases. As shown in Fig. 3, there are three key components: spatial reduction attention, dual-path attention, and feature enhancement module in TransDD.

Spatial reduction attention (SRA). The self-attention mechanism which has a larger range of receptive fields than CNN can capture variances in appearance, location, and scale of the lesion regions in CXR images. The complexity of global self-attention depends on the number of patches (hw). In detail, the computational complexity is proportional to the square of the number of patches. To reduce the complexity of the global self-attention, we perform average pooling followed by linear projection to reduce the resolution of the feature and get a matrix $R \in \mathbb{R}^{\frac{hw}{r^2} \times hw}$, as shown in Fig. 3(a). Here, r is set to 4. After that, the R is set as an intermediate for the similarity comparison, instead of directly multiplying by the transposition of $Q \in \mathbb{R}^{hw \times d}$ and $K \in \mathbb{R}^{hw \times d}$. Formally, the following equations can describe the calculation process. Here, R is set as an intermediate for the similarity comparison. Q, K, and V, which are obtained through three linear projections, represent the query, key, and value, respectively. RK means the similarity between R and K, whereas QR denotes the similarity between Q and R.

$$RK = Softmax(\frac{R \bullet K^T}{\sqrt{d}}) \tag{1}$$

$$QR = Softmax(\frac{Q \bullet R^T}{\sqrt{d}}) \tag{2}$$

$$Y = QR \bullet (RK \bullet V) \tag{3}$$

Dual-path attention (DPA). The architecture of dual-path attention is similar to spatial attention. From Fig. 3, it can be observed the difference between them in detail. (1) The inputs of SRA all come from the same input, while the L comes from the label encoder. (2) We use a feature enhancement module to extract the local feature information. (3) After conducting two similarity comparisons, DPA is capable of establishing a stronger correlation between feature maps and their respective labels compared to conventional cross-attention. The following equations can describe the calculation process. Here, L is from the label encoder, whereas Q, K, and V are from feature maps which are yielded by three linear projections. LK means the similarity between L and K, whereas QL denotes the similarity between Q and L.

$$LK = Softmax(\frac{L \bullet K^T}{\sqrt{d}}) \tag{4}$$

$$QL = Softmax(\frac{Q \bullet L^T}{\sqrt{d}}) \tag{5}$$

$$Y = QL \bullet (LK \bullet V) \tag{6}$$

Feature enhancement module (FEM). As shown in the green dashed rectangular box in Fig. 2, FEM is designed to enhance the expressiveness of local features, which leverage simple convolution and pooling operations to effectively reduce model parameters and computation. The core of FEM is to use two 1×1 convolution operations to interact with features among different channels. Specifically, it can be expressed as Eq. (7). Here, BN is Batch normalization and $ReLu$ represents an activation function.

$$FEM(f) = BN(Conv(ReLu(BN(Conv(f))))) \tag{7}$$

In addition, the feed-forward layer is a multilayer perceptron (MLP) that is utilized to select the features. M_1 and M_2 denote the number of label decoders and feature decoders, respectively.

3.2.2. Classification attention block

The pipeline of classification attention block (CAB) is illustrated in Fig. 4, while the pseudo-code is described in Fig. 5. After obtaining two classification scores from the feature output $F_O \in \mathbb{R}^{hw \times d}$ and label output $L_O \in \mathbb{R}^{N \times d}$, CAB is designed to balance the label score $S_{L_O} \in \mathbb{R}^d$ and feature score $S_{F_O} \in \mathbb{R}^d$. Here, the label score is yielded by calculating the mean score of the label output along the dimension of the column. Meanwhile, we identify the maximum value across all spatial locations for each category. It focuses our attention on classifying scores at different locations for different disease categories. This particular mechanism can be viewed as a class-specific attention approach. This attention mechanism is intuitively very useful for multi-label recognition, particularly in scenarios where there are objects from numerous classes and with varying sizes. Then, two linear projection layers (label head and feature head) are utilized to project the dimension of the label score and feature score from d to c. Finally, la is set as a hyperparameter to balance them. CAB provides our model with the ability to efficiently pinpoint and evaluate the classification scores for various object

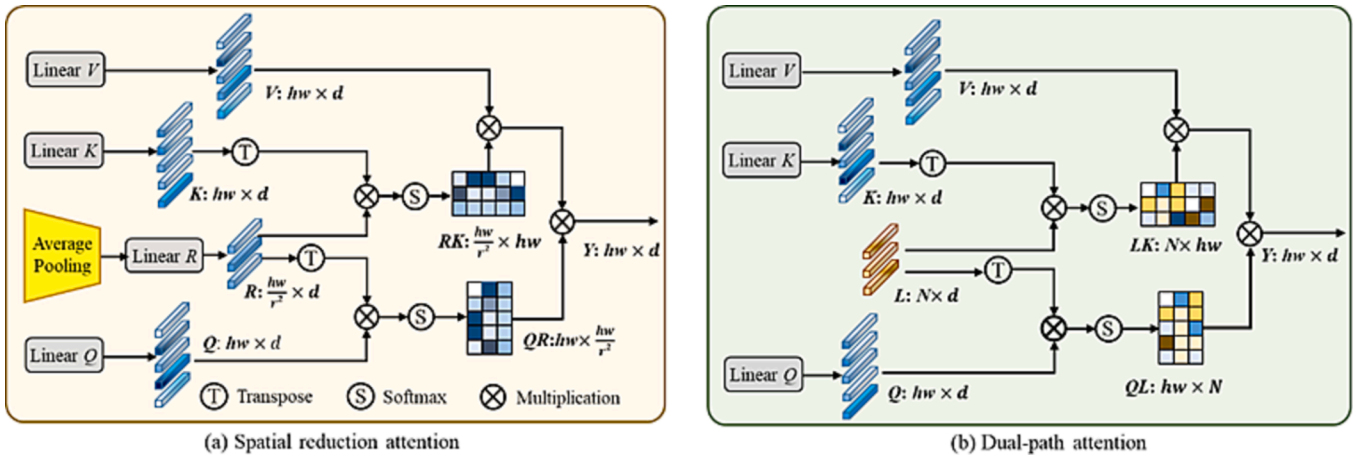

Fig. 3. The architecture of the proposed spatial reduction attention (a) and dual-path attention (b) in the feature decoder. The designed SRA can capture variances in appearance, location, and scale of the lesion regions in CXR images. And, the DPA is used to establish the connection between local discriminative features and the corresponding label.

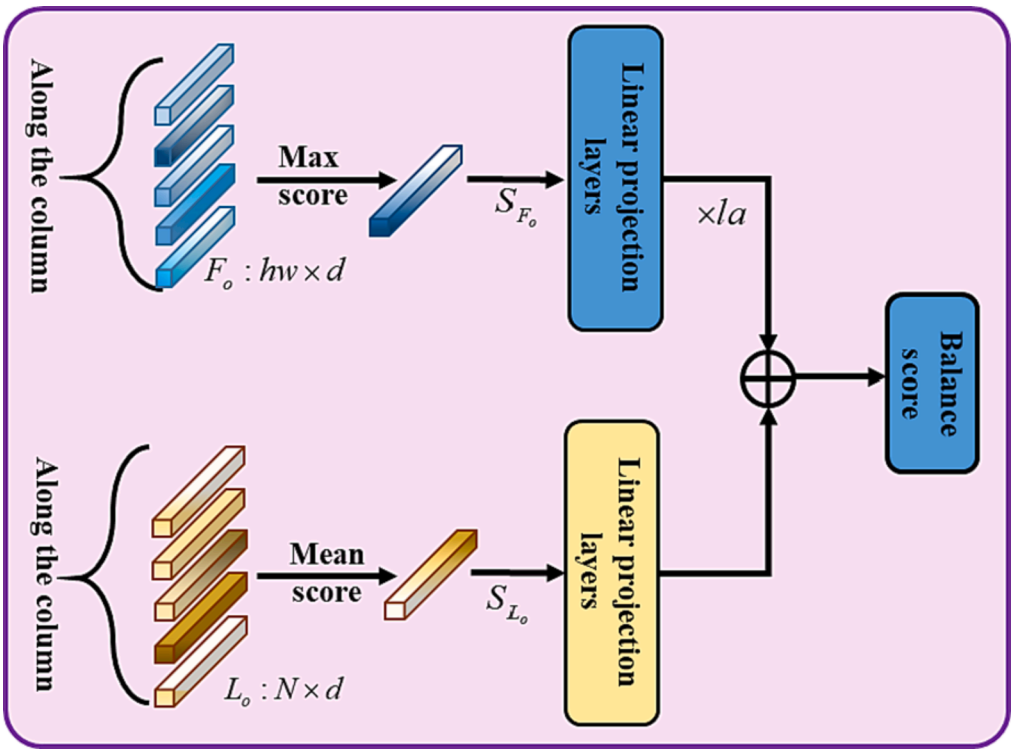

Fig. 4. The pipeline of classification attention block (CAB).

#Pseudo code of classification attention block

```
class CAB(nn.Module):
#la: a hyperparameter
#c: number of  categories in dataset
#label_output, shape: (N, d)
#feature_output, shape: (hw, d)
   def __init__(self, d, c, la):
      super().__init__()
      self.la=la
      self.label_head= nn.Linear(d, c)
      self.feature_head=nn.Linear(d, c)

   def forward(self, label_output, feature_output):
      #label_score, shape: (c)
      label_score=self.label_head(label_output.mean(dim=1))
      #feature_score, shape: (c)
      feature_score=self.feature_head(torch.max(feature_output, dim=1)[0] )
      balance_score=label_score+self.la*feature_score
      return balance_score
```

Fig. 5. The pseudo-code of classification attention block (CAB).

categories across varying spatial locations.

3.3. Training and evaluation strategy

Multi-label classification of thoracic diseases. For the ChestX-ray14 dataset, we followed the official split that is publicly available on the NIH website which separates the dataset into a training set of 86,524 images and a testing set of 25,596 images. The distribution of training and testing images overall categories in the ChestX-ray14 dataset is given in Fig. 6, which emphasizes that the datasets are highly imbalanced and each image may have one or more types. In addition, we strictly follow the official patient-wise split standards provided by [8] that the SYSU dataset is randomly assigned for training (80 %), validation (10 %), and testing (10 %), as shown in Fig. 7. Especially, SYSU-PE dataset is solely leveraged for external validation.

Multi-class classification of thoracic diseases. Following the [8], we adopted two-classification (COVID-19, and Non-COVID-19) and three-class classifications (normal, viral pneumonia, and other

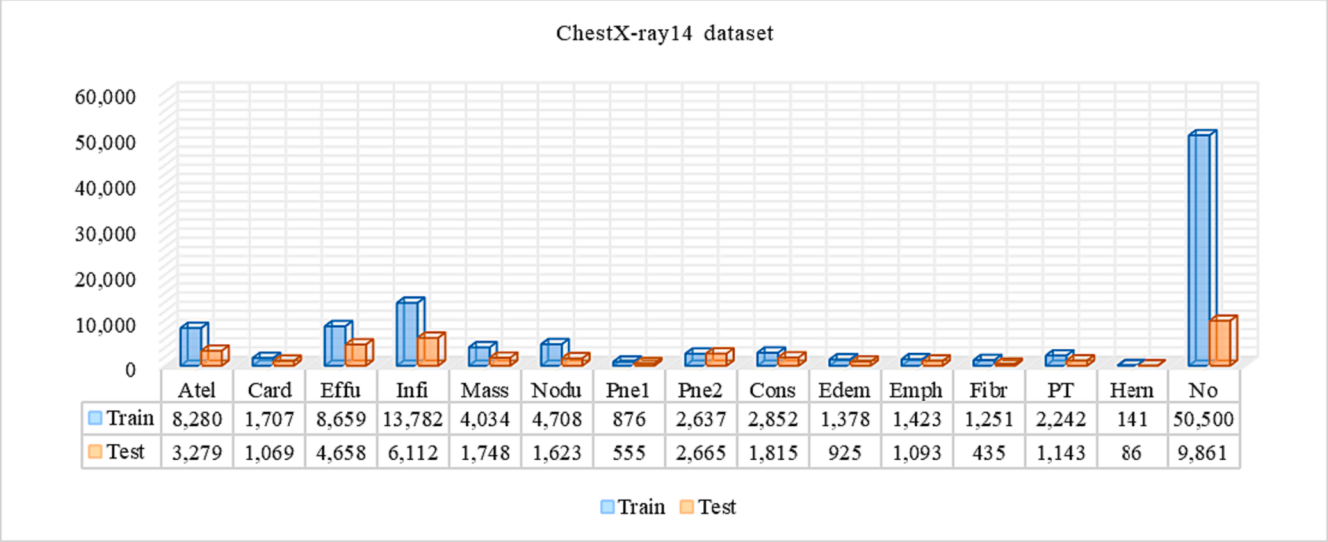

Fig. 6. Distribution of training and testing images overall categories in the ChestX-ray14 dataset. The 15 categories are Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, Hernia, and No Finding, respectively.

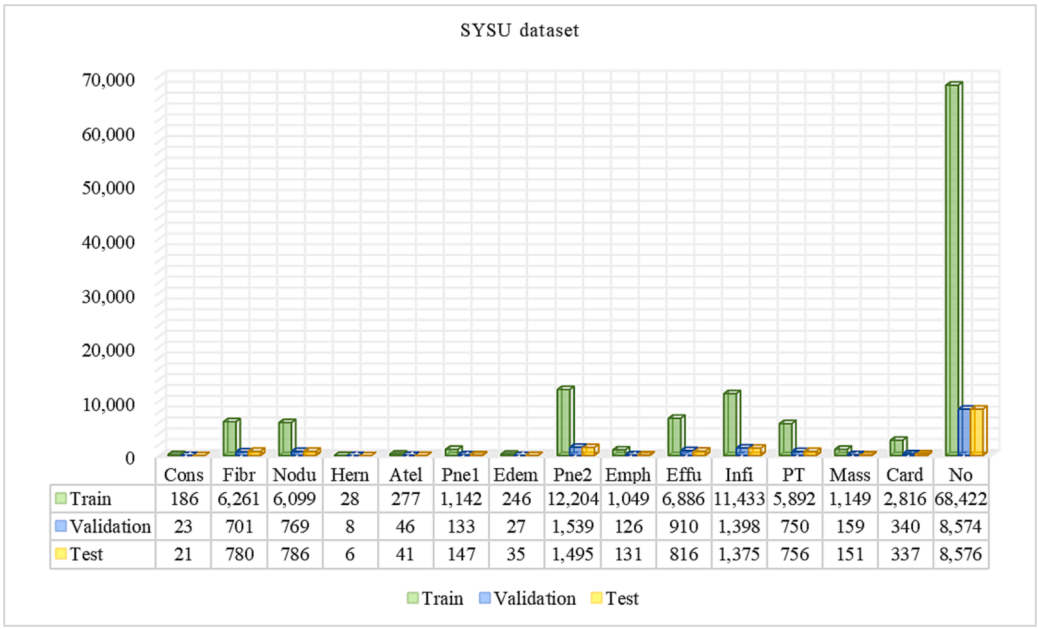

Fig. 7. The number of 15 categories for training, validation, and testing in the SYSU dataset. *The 15 categories are Consolidation, Fibrosis, Nodule, Hernia, Atelectasis, Pneumothorax, Edema, Pneumonia, Emphysema, Effusion, Infiltration, Pleural_thickening, Mass, Cardiomegaly, and No_finding, respectively.

pneumonia) on the CC-CXRI-P dataset. As shown in Table 1, the dataset is also assigned for training (80 %), validation (10 %), and testing (10 %).

The original image size in the ChestX-ray14 dataset is 1024 × 1024, whereas the size in other datasets is various. To streamline computational complexity and ensure uniform standards for experimental purposes, all images are resized to 384 × 384. Then, the CXR images were enhanced by center crop and random horizontal flips. Finally, we trained our framework end-to-end while the feature encoder has been pre-trained on the ImageNet-1 k dataset [35]. The CXR images were also resized to 384 × 384 for validation and testing. Here, each image is labeled with $y = \{y_1, y_2, \cdots y_N\}$, and N is the number of categories in the dataset.

Moreover, asymmetric loss [36] is employed to deal with the problem of class imbalance, which is a variant of focal loss with different γ values for positive and negative values. $P = \{P_1, P_2, \cdots P_N\}$ denotes the output of our framework. Then the asymmetric loss (ASL) to calculate the loss for each CXR image is shown in Eq. (8). Followed by [36], we set $\gamma^+ = 0$ and $\gamma^- = 4$ in our experiments as default without tuning.

Table 1
The number of CXR images for training, validation and testing in the CC-CXRI-P dataset.

	Normal	Viral pneumonia		Other pneumonia	Total
		COVID-19	Non-COVID-19		
Training	2,904	489	1,328	1,617	6,338
Validation	363	62	166	202	793
Testing	362	61	165	202	790

$$ASL = \frac{1}{N}\sum_{N=1}^{N}\begin{cases}(1-P_N)^{\gamma^+}log(P_N), y_N = 1\\(P_N)^{\gamma^-}log(1-P_N), y_N = 0\end{cases} \quad (8)$$

3.4. Implementation details

The proposed TransDD framework is implemented by using Python 3.7 and Pytorch 1.7.0 that PyCharm as our IDE while running on 2 Nvidia 3090 GPUs with 48 GB memory. During the training period, our model utilizes a mini-batch size of 64 and an initial learning rate set at $1e-4$, which gradually decreases by the cosine schedule [37] over the course of 20 epochs. Furthermore, we employ the prominent AdamW [38] optimizer with a momentum of 0.9 and weight decay of $1e-3$ to optimize the back propagation process.

4. Results

The proposed TransDD framework was evaluated on the ChestX-ray14, SYSU and SYSU-PE datasets for multi-label classification of thoracic diseases, while validating the ability of multi-class classification of thoracic diseases on the CC-CXRI-P dataset.

4.1. Evaluation for multi-label classification of thoracic diseases

We first perform extensive experiments on the ChestX-ray14, SYSU, and SYSU-PE datasets to evaluate the ability of our proposed TransDD framework for multi-label classification of thoracic diseases. Since this is a multi-label classification task instead of a single-label classification task and the datasets are extremely imbalance, the area under the receiver operating characteristic curve (AUC) [42] is a more reasonable performance metric which has been employed in most related work [1,3,7,16] than other metrics such as accuracy. The comparative AUC results of our TransDD-PVT framework and the state-of-the-art (SOAT) previous works on the ChestX-ray14 dataset are presented in Table 2. Moreover, we also compare the proposed TransDD framework to the CNN-based models (i.e., ResNet101 [14], and DenseNet121 [15]) and Transformer-based models (Swin-B [20] and PVTv2-B4 [41]) on the SYSU dataset, as shown in Table 3.

Meanwhile, the SYSU-PE dataset is employed for external validation to verify the generalization of the model. The receiver operating characteristic curves (ROC) of PVTv2-B4 model and the proposed TransDD-PVT framework over the 4 pathologies on the SYSU-PE dataset are illustrated in Fig. 8.

Based on the above results, we can gain the following observations: (1) The proposed TransDD-PVT framework achieved the highest mean AUC score of 83.1 % for all 14 pathologies on ChestX-ray14, while yielding the top performance for more than half of pathologies. (2) Compared to other existing works, our TransDD-PVT utilizes the specially designed SRA to effectively capture variations in the appearance, location, and scale of lesion regions in CXR images. Moreover, the DPA is used to establish the connection between local discriminative features and the corresponding label. Therefore, our designed dual-path decoder can effectively improve the performance of the backbone. For instance, the mean AUC score of Our TransDD-ResNet is improved by 1.1 % than ResNet101. (3) Note that the SYSU-PE dataset which is consists of additional patients who underwent a routine annual physician examination has not used in training. From Fig. 8, it can be seen that our TransDD-PVT has good generalization performance and can distinctly improve the robustness and AUC score of pathologies, i.e., Fibrosis (93.2 % vs. 93.5 %), Nodule (85.8 % vs. 87.3 %), Pneumothorax (94.6 % vs. 94.9 %) and Pleural thickening (91.4 % vs. 92.0 %). (4) Overall, these comparative results demonstrate the effectiveness of our TransDD framework for multi-label classification of thoracic diseases.

4.2. Evaluation for multi-class classification of thoracic diseases

We further evaluate the performance of our TransDD framework for multiclass classification of thoracic diseases on the CC-CXRI-P dataset through three-class classification (normal, viral pneumonia, and other pneumonia) and two-classification (COVID-19, Non-COVID-19). To quantitatively compare the CNN-based models (i.e., ResNet101 [14] and DenseNet121 [15]) and Transformer-based models (Swin-B [20] and PVTv2-B4 [41]) in identifying viral pneumonia, we calculated the test accuracy (ACC), sensitivity (SEN), and precision (PRE) of each infection type. The comparative confusion matrices of TransDD framework and other SOAT backbones are illustrated in Fig. 9, while the average ACC, SEN, and PRE are presented in Table 4. Moreover, the performance of Swin-B and our TransDD-Swin on two-classification (COVID-19, Non-COVID-19) are shown in Fig. 10. Based on these various experimental results, we can gain some new observations as follows: (1) By taking advantage of the ability of our TransDD framework, our methods outperform all other contrastive methods on all metrics with a large margin in discriminating between viral pneumonia, other types of pneumonia and the absence of pneumonia from CXR images. For example, the ACC score of TransDD-ResNet and TransDD-DenseNet increased by 1.01 % and 1.27 % comparing the baselines, respectively. Meanwhile, the SEN score of TransDD-Swin and TransDD-PVT are more 1.49 % and 0.84 % than the baselines. (2) According to the Fig. 10, it is clear see that our TransDD-PVT also achieve better metrics in discriminating between other viral pneumonia and COVID-19 pneumonia from CXR images. (3) Sufficient experiments show that our TransDD framework has a great ability of identifying viral pneumonia.

4.3. Visualization of attention heat maps

Based on the above objective analysis, a gradient-weighted class activation map (Grad-CAM) [43] is utilized to generate the heatmaps of the CXR images which can approximately visualize the indicative attention areas, as shown in Fig. 11. Note that we do not add any bounding boxes for training or testing. The lower response is demonstrated in blue while the higher is highlighted in red. Compared to the PVTv2-B4, we can find that our TransDD-PVT framework can locate of lesion more precisely. For example, the heatmaps with Atelectasis and Pneumonia generated by our TransDD-PVT can produce smaller red responses than PVTv2-B4 for the small scale of the lesion regions. Meanwhile, the heatmaps with Cardiomegaly and Infiltrate generated

Table 2
Comparison results of comparative methods on the ChestX-ray14 dataset. We illustrate the AUC score (%) of each disease pathology and the average AUC scores (%) across the 14 classes. Significantly, the highest scores are shown in bold.

Method	Atel	Card	Effu	Infi	Mass	Nodu	Pne1	Pne2	Cons	Edem	Emph	Fibr	PT	Hern	Mean
U-DCNN [7]	70.0	81.0	75.9	66.1	69.3	66.9	65.8	79.9	70.3	80.5	83.3	78.6	68.4	87.2	74.5
Thorax-net [27]	75.1	87.1	81.2	68.1	79.9	71.5	69.4	82.5	74.2	83.5	84.3	80.4	74.6	90.2	78.8
CheXNet [26]	76.9	88.5	82.5	69.4	82.4	75.9	71.5	85.2	74.5	84.2	90.6	82.1	76.6	90.1	80.7
CRAL [16]	78.1	88.0	82.9	70.2	83.4	77.3	72.9	85.7	75.4	85.0	90.8	83.0	77.8	91.7	81.6
DualCheXNet [39]	78.4	88.8	83.1	70.5	**83.8**	79.6	72.7	87.6	74.6	85.2	94.2	83.7	79.6	91.2	82.3
LLAGnet [40]	78.3	88.5	83.4	70.3	84.1	79.0	72.9	87.7	75.4	85.1	93.9	83.2	79.8	91.6	82.4
A^3 Net [1]	77.9	**89.5**	83.6	71.0	83.4	77.7	73.7	87.8	**75.9**	85.5	93.3	83.8	79.1	**93.8**	82.6
TransDD-PVT (Ours)	**79.1**	88.5	**84.2**	**71.5**	83.7	**80.3**	**74.5**	**88.5**	75.3	**85.9**	**94.4**	**84.9**	**80.3**	92.4	**83.1**

Table 3
Comparison results of comparative methods on SYSU dataset. We illustrate the AUC score (%) of each disease pathology and the average AUC scores (%) across the 14 classes. Significantly, the highest scores are shown in bold.

	Cons	Fibr	Nodu	Hern	Atel	Pne1	Edem	Pne2	Emph	Effu	Infi	PT	Mass	Card	Mean
ResNet101 [14]	97.1	89.0	83.1	88.1	94.9	**98.2**	98.2	91.5	95.9	97.4	93.5	89.9	94.3	96.3	93.4
TransDD-ResNet	**97.2**	**89.4**	**83.6**	**96.8**	**95.9**	97.9	**98.5**	**92.0**	95.9	**97.7**	**93.8**	**90.5**	**95.0**	**96.4**	**94.3**
DenseNet121 [15]	97.1	89.6	83.6	90.3	94.0	98.2	98.7	92.1	96.2	97.6	93.9	90.8	**94.6**	96.2	93.8
TransDD-DenseNet	**97.9**	**89.9**	**84.0**	**94.5**	**95.5**	**98.3**	**98.8**	**92.4**	96.2	**97.8**	**94.2**	**91.1**	94.5	**96.7**	**94.4**
Swin-B [20]	97.6	89.8	82.2	90.9	95.8	98.2	98.8	92.4	96.0	97.7	94.2	90.8	95.1	96.6	94.0
TransDD-Swin	**97.9**	**90.3**	**83.4**	**92.1**	95.7	**98.4**	**98.9**	92.3	**97.3**	**98.2**	**95.9**	90.1	**95.6**	96.3	**94.5**
PVTv2-B4 [41]	97.0	90.3	84.3	91.3	95.3	98.4	98.4	92.5	96.6	**97.6**	94.2	90.3	94.7	96.4	94.1
TransDD-PVT	**97.7**	**90.6**	**84.5**	**91.6**	**96.4**	**98.6**	**98.9**	**93.1**	**97.2**	97.5	**94.6**	**91.4**	**95.0**	**96.8**	**94.6**

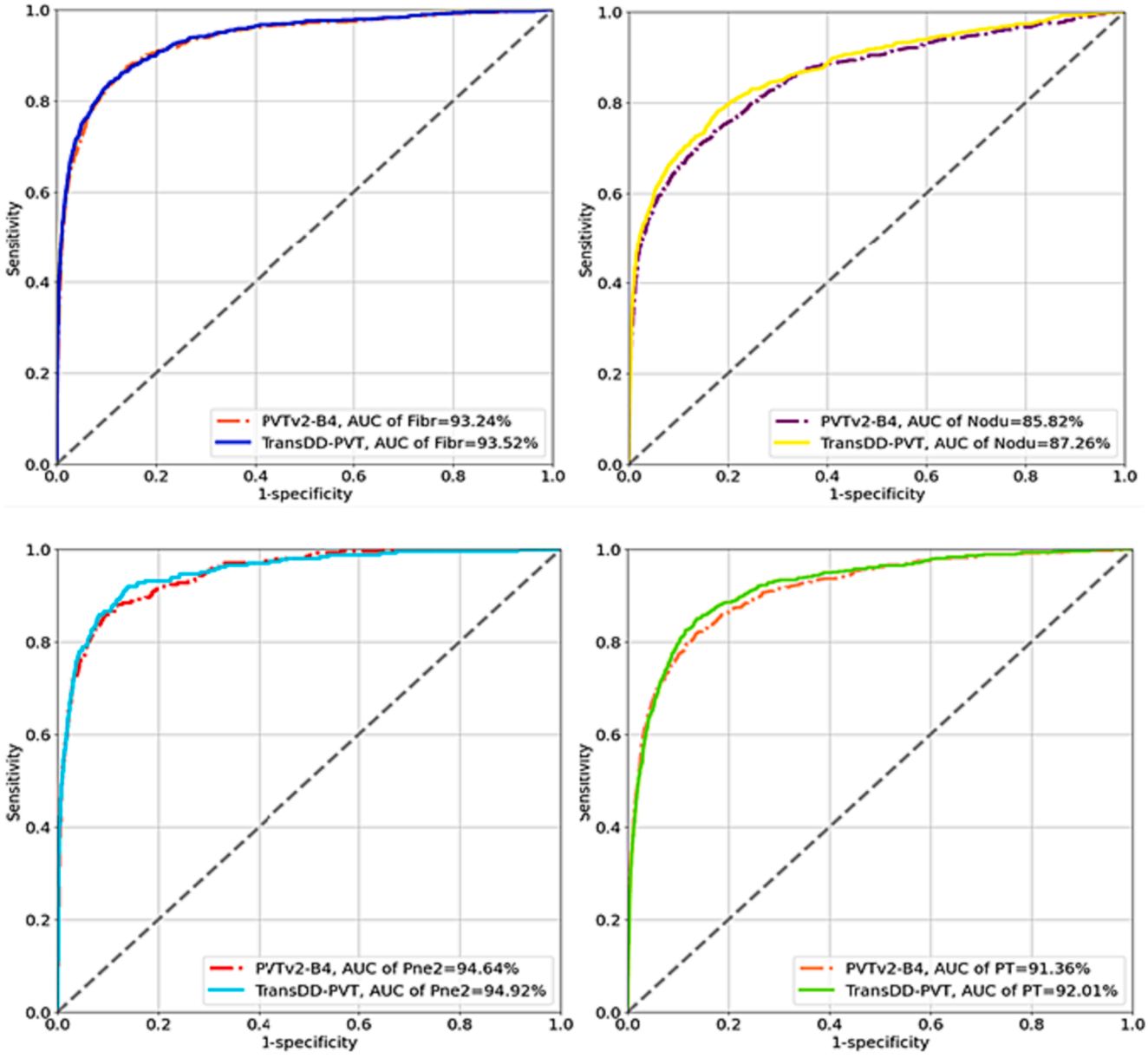

Fig. 8. ROC curves of the proposed TransDD-PVT and PVTv2-B4 model on SYSU-PE dataset over 4 pathologies for external validation.

by PVTv2-B4 are fuzzy while our TransDD-PVT can yield clear and precise heatmaps. Moreover, the PVTv2-B4 mislocates the lesion regions with Mass and Effusion. Therefore, we can find that the proposed dual-path decoder block is not only applicable for disease classification but also has a powerful ability for localizing lesion regions with scale variance and different locations of the lung field.

4.4. *Visualizing the distribution of feature representations*

In addition, we also perform t-Distributed Stochastic Neighbor Embedding (t-SNE) [44] which is a statistical method for visualizing high-dimensional data by giving each data point a location in a two or three-dimensional map. The distribution of features with different

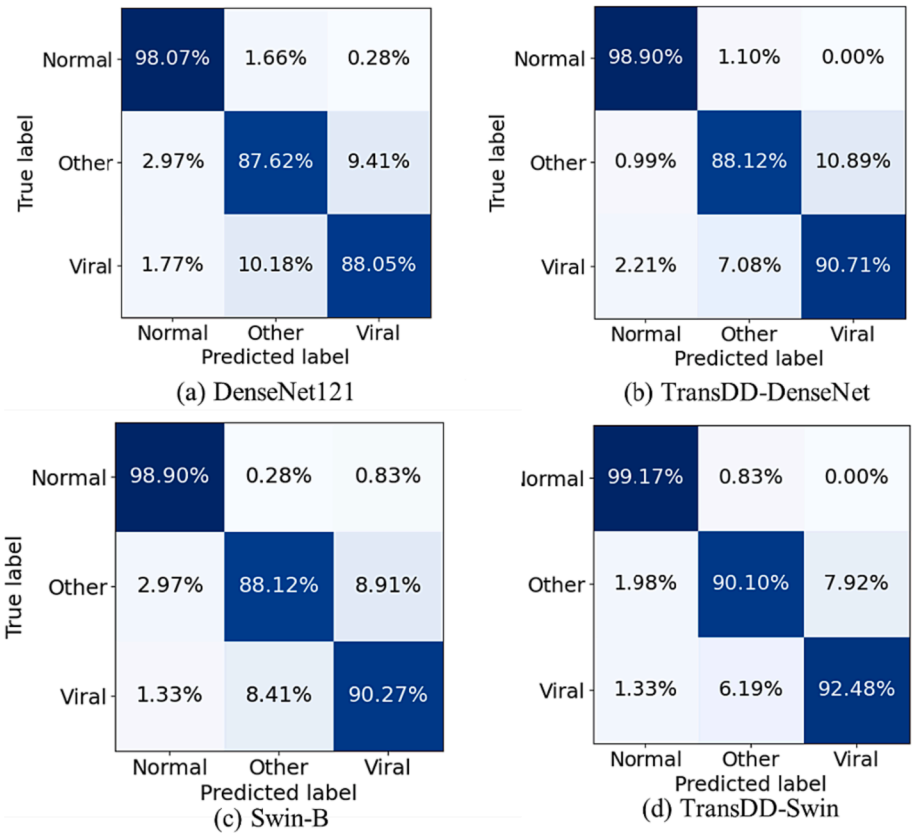

Fig. 9. Comparison of the performance of our TransDD framework to known state-of-the-art models in discriminating between viral pneumonia, other types of pneumonia, and the absence of pneumonia.

Table 4
The performance of comparative backbones and our TransDD framework in identifying viral pneumonia (three-class classification).

	ACC (%)	SEN (%)	PRE (%)
ResNet101 [14]	92.15	90.83	90.75
TransDD-ResNet(ours)	**93.16**	**92.09**	**92.01**
DenseNet121 [15]	92.53	91.25	91.35
TransDD-DenseNet (ours)	**93.80**	**92.57**	**92.76**
Swin-B [20]	93.67	92.43	92.70
TransDD-Swin (ours)	**94.94**	**93.92**	**94.14**
PVTv2-B4 [41]	93.79	92.47	92.83
TransDD-PVT (ours)	**94.31**	**93.31**	**93.21**

thoracic diseases extracted from PVTv2-B4 and our TransDD-PVT is illustrated in Fig. 12. In Fig. 12(a), it is clear that our TransDD-PVT can gather most of the test samples in the same compact region and display clear boundaries between categories. By contrast, the distribution learned through the PVTv2-B4 appears more confusing, especially for Mass and Fibrosis. Meanwhile, we draw the same observation in Fig. 12 (b), especially for viral Pneumonia and other pneumonia. Based on the above analysis, it can further prove that our TransDD-PVT has a powerful ability for guiding and recalibrating feature learning with semantic consistency.

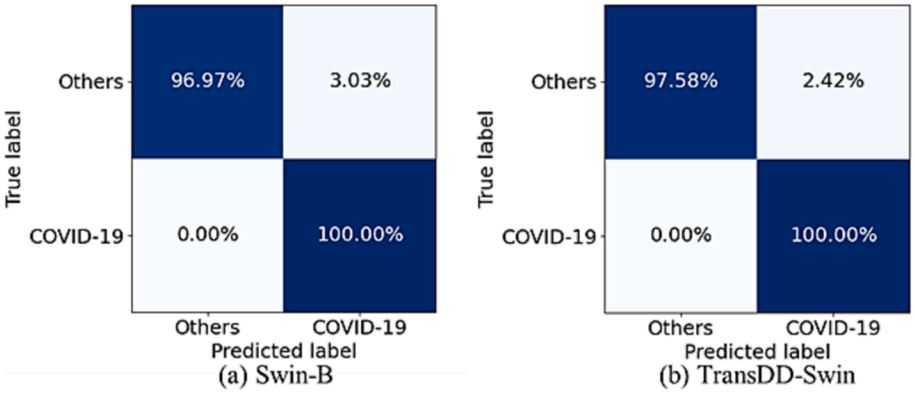

Fig. 10. Comparison of the performance of the Swin-B to our TransDD-Swin in identifying COVID-19 pneumonia from CXR images.

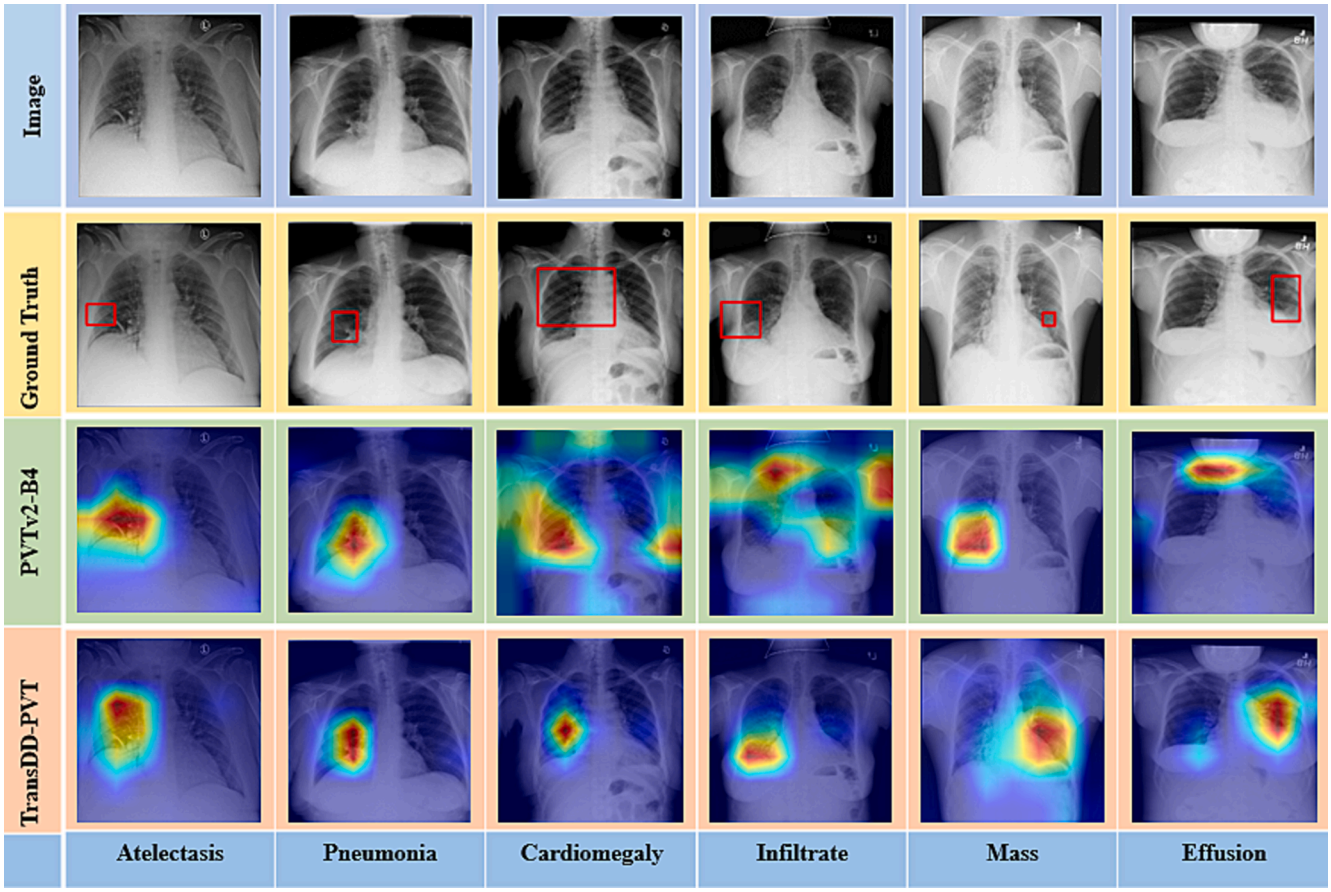

Fig. 11. Localization of lesion regions with PVTv2-B4 and the proposed TransDD-PVT framework. The original images are shown in the first line and the manual lesion regions provided by the official version are annotated with red bounding boxes in the second line. The heatmaps generated by PVTv2-B4 and our proposed TransDD-PVT are illustrated in the third and the fourth line, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5. Discussion

5.1. Combination of label decoder and feature decoder

In this paper, we proposed a dual-path decoder block which consists of the label decoder and the feature decoder. Naturally, there are many different combinations of label decoder and feature decoder. To appropriateness of our combination, we compared the performance of various methods on the CC-CXRI-P dataset. The mean ACCs, SENs, and PREs of the proposed TransDDDenseNet framework are shown in Table 5. It can be seen that the dual-path decoder block with two label decoders and one feature decoder is more beneficial in boosting diagnostic performance.

5.2. Selecting the dimension of the learnable label

Moreover, a learnable label $L \in \mathbb{R}^{N\times d}$ is set to establish category-related features from the spatial features, where N is the number of categories and d is the dimension of the learnable label. To verify the influence of using different values of d, we performed the ablation experiment on the CC-CXRI-P dataset. In particular, TransDD-DenseNet and TransDD-PVT with two label decoders and one feature decoder are utilized to evaluate the influence of the dimension of the learnable. From Fig. 13, it can be seen that the dimension of the learnable label is related to the dimension of the feature extracted by the backbone. For example, the dimension of the learnable label of TransDD-DenseNet is set as 1024 as same as the dimension of the feature that can achieve better performance. Conversely, the dimension of the learnable label of TransDD-PVT is set as 512.

5.3. Impact of spatial reduction attention and dual-path attention

In addition, we also investigate the impact of the proposed spatial reduction attention (SRA) and dual-path attention (DPA) on the CC-CXRI-P dataset. Here, conventional self-attention (SA) and cross-attention (CA) are used for comparison. From Table 6, we can find that the proposed SRA and DPA outperform the conventional attention mechanism. In detail, the DPA can achieve more 0.59 % ACC than CA, while the SRA attains more 0.53 % than SA. Moreover, the ACC increases by 1.11 % when using both SRA and DPA.

5.4. Selection of classification scores in classification attention block

The feature output $F_O \in \mathbb{R}^{hw\times d}$ and label output $L_O \in \mathbb{R}^{N\times d}$ are obtained after the dual-path decoder block. Then, we consider selecting which one to use for classification. Therefore, we utilized TransDD-DenseNet with two label decoders and one feature decoder as our pipeline and compared the performance on the CC-CXRI-P dataset. The results of ACCs and SENs are shown in Fig. 14. Method A only uses feature output while Method B utilizes label output. Meanwhile, Method C and Method D use both the feature output and label out for classification. The difference between them is that the feature output in Method C performed the average pooling while performing the max pooling in Method D. On the basis of Fig. 14, we can find that Method D yields the highest ACC and SEN for thoracic diseases classification. The proposed CAB module can find the maximum value among all spatial locations for

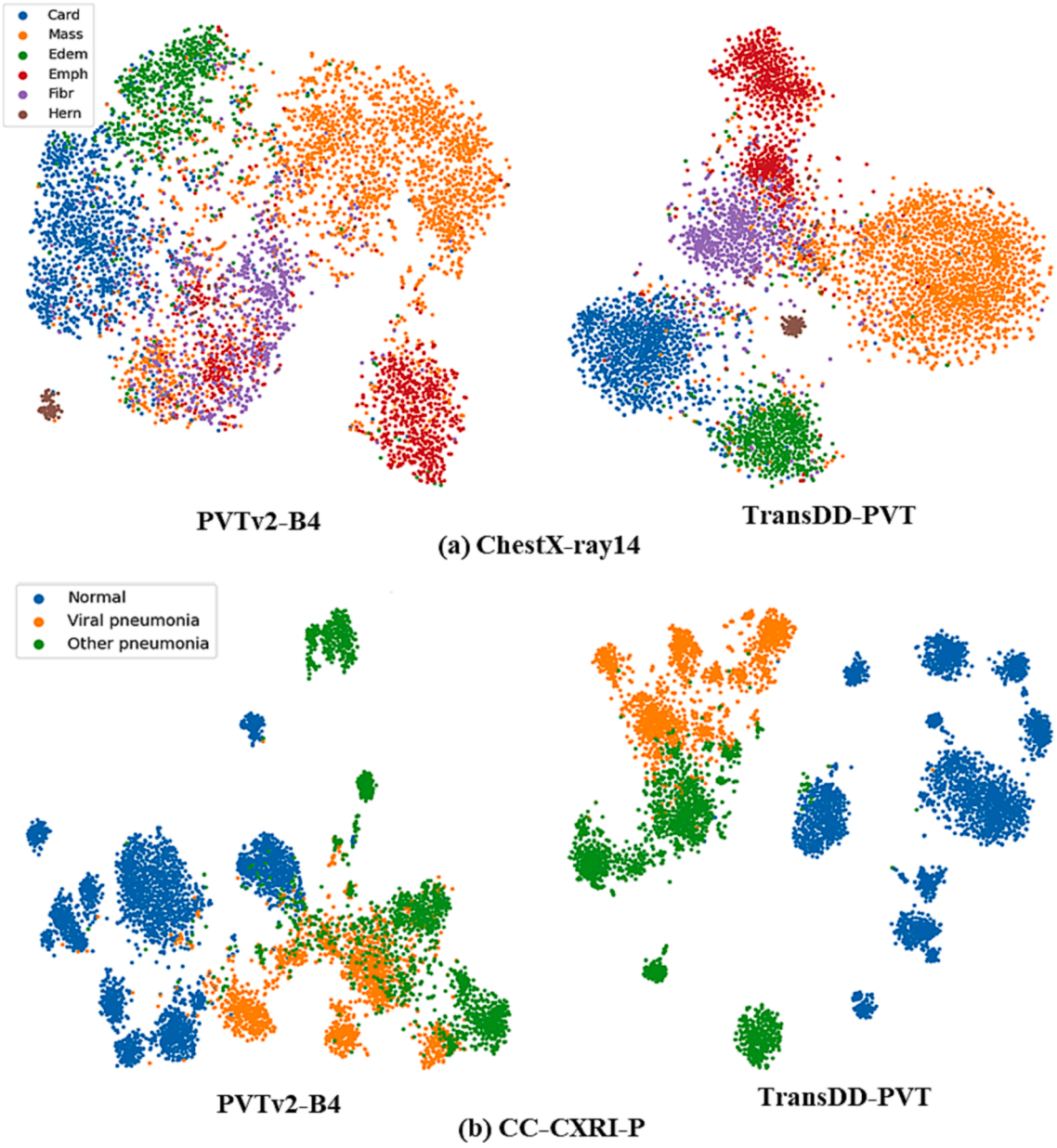

Fig 12. Visualization of the feature representations learned by PVTv2-B4 and the proposed TransDD-PVT framework on the ChestX-ray14 dataset (a) and the CC-CXRI-P dataset (b), respectively. The single-label CXR image is marked with different colors.

Table 5
The mean ACC, SEN and PRE of the proposed TransDD-DenseNet framework when using different combinations of label decoder and feature decoder. M_1 and M_2 denote the number of label decoders and feature decoders, respectively.

M_1	M_2	ACC(%)	SEN(%)	PRE(%)
0	0	92.53	91.25	91.35
1	0	92.15	90.55	90.95
1	1	92.53	91.41	91.26
1	2	93.16	91.77	92.12
2	1	**93.56**	**92.13**	**92.86**
2	2	93.40	92.08	92.74

each category feature score via max pooling. Furthermore, *la* is set as a hyperparameter to balance them and we find that the proposed TransDD-DenseNet can produce a more accurate diagnosis when *la* is set to 0.8.

5.5. Merits and limitations

The proposed TransDD framework has a distinct advantage over other models in optimizing visual feature embedding and label embedding. Extensive experiments demonstrate the effectiveness of our framework for improving the performance of thoracic disease classification. However, the proposed method still reveals some limitations. First, the disease labels in the ChestX-ray14 are noisy, since they were mined from the radiological reports using NLP techniques. We used all of the labels without discriminating against them. Second, we had not considered additional information provided by the datasets, such as patient age, gender, medical history, and clinical symptoms views. In future work, we will further consider these limitations and propose the corresponding methods. In addition, transfer learning and few-shot learning may be viable directions for future research opportunities and potential trends in the classification of thoracic diseases. Thoracic diseases classification often suffers from limited labeled data. Transfer learning and few-shot learning techniques can be explored to overcome

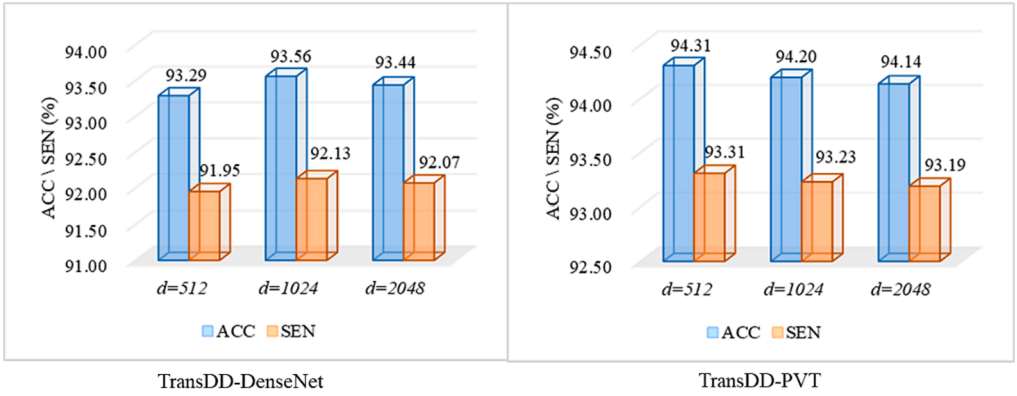

Fig. 13. Comparison of our TransDD framework with different dimensions of the learnable label.

Table 6
Ablation study on attention mechanism.

Attention mechanism	ACC(%)	SEN(%)
SA + CA	92.45	91.15
SRA + CA	92.98	91.89
SA + DPA	93.04	91.92
SRA + DPA (Ours)	**93.56**	**92.15**

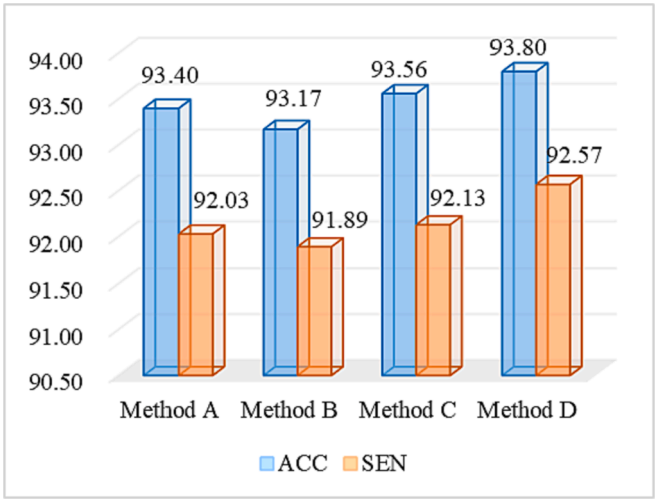

Fig. 14. Comparison of our classification attention block with different classification scores. Method A only uses feature output while Method B utilizes label output. Meanwhile, Method C performed the average pooling, and Method D used the max pooling for feature output.

this challenge by leveraging pre-trained models and transferring knowledge from related tasks.

6. Conclusion

In this paper, we innovatively introduce a learnable label embedding as queries to detect and match class-related features from the feature maps, and then computed by a novel Transformer-based dual-path decoder (TransDD). The proposed TransDD can serve as a plug-and-play structure to improve the thoracic diseases classification performance of both CNNs and recent Transformer-based backbones. To capture variances in appearance, location, and scale of the lesion regions and reduce the complexity of global self-attention, we design spatial reduction attention. And dual-path attention is designed to connect the explicit correlation between the features and labels. Furthermore, we utilize a classification attention block to balance two classification scores based on feature output and label output, respectively. Extensive experiments conducted on several datasets demonstrate the powerful ability of our TransDD to localize lesion regions with varying scales and different locations within the lung field. This capability brings a significant boost on the comparative backbones. In future work, we plan to analyze the local discriminative diseased features and corresponding labels in greater detail. We also intend to utilize semi-supervised learning to rely less on noisy labels.

CRediT authorship contribution statement

Xiaoben Jiang: Conceptualization, Methodology, Software, Validation, Writing – original draft, Visualization, Investigation. **Yu Zhu:** Supervision, Writing – review & editing, Project administration. **Yatong Liu:** Resources, Data curation. **Gan Cai:** Formal analysis, Data curation. **Hao Fang:** Project administration, Funding acquisition.

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.

Data availability

Data will be made available on request.

Acknowledgment

The authors greatly appreciate the financial support of National Scientific Foundation of China (81971863, 82272230), Shanghai Natural Science Foundation (22ZR1444700), Shanghai Shenkang project for transformation for scientific production (SHDC2022CRD049).

References

[1] H. Wang, S. Wang, Z. Qin, Y. Zhang, R. Li, Y. Xia, Triple attention learning for classification of 14 thoracic diseases using chest radiography, Med. Image Anal. 67 (2021) 101846.
[2] M. Hiles, A.-M. Culpan, C. Watts, T. Munyombwe, S. Wolstenhulme, Neonatal respiratory distress syndrome: chest X-ray or lung ultrasound? A Systematic Review, Ultrasound Med Biol. 25 (2) (2017) 80–91.
[3] B. Chen, Z. Zhang, Y. Li, L.u. Guangming, D. Zhang, Multi-Label Chest X-ray Image Classification via Semantic Similarity Graph Embedding, IEEE Transactions on Circuits Systems for Video Technology. 32 (4) (2021) 2455–2468.
[4] D.E. Litmanovich, M. Chung, R.R. Kirkbride, G. Kicska, J.P. Kanne, Review of chest radiograph findings of COVID-19 pneumonia and suggested reporting language, J. Thorac. Imaging 35 (6) (2020) 354–360.
[5] O.h. Yujin, Sangjoon Park, Jong Chul Ye, Deep learning covid-19 features on cxr using limited training data sets, IEEE Trans Med Imaging. 39 (8) (2020) 2688–2700.

[6] Zhi Zhen Qin, Melissa S Sander, Bishwa Rai, Collins N Titahong, Santat Sudrungrot, Sylvain N Laah, Lal Mani Adhikari, E Jane Carter, Lekha Puri, Andrew J Codlin, Using Artificial Intelligence to Read Chest Radiographs for Tuberculosis Detection: A Multi-Site Evaluation of the Diagnostic Accuracy of Three Deep Learning Systems, Scientific Reports. 9 (1) (2019) 1–10.

[7] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, Ronald M Summers: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2097-2106.

[8] G. Wang, X. Liu, J. Shen, C. Wang, Z. Li, L. Ye, W.u. Xingwang, T. Chen, K. Wang, X. Zhang, A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images, Nat. Biomed. Eng. 5 (6) (2021) 509–521.

[9] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[10] Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE. 86 (11) (1998) 2278-2324.

[11] Alex Krizhevsky, I. Sutskever, G. Hinton: ImageNet classification with deep convolutional neural networks. In: Conference on Neural Information Processing Systems (NIPS), 2012, pp. 1097-1105.

[12] Karen Simonyan, Andrew Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv: 1409.1556.

[13] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich: Going deeper with convolutions. In: Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1-9.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun: Deep residual learning for image recognition. In: Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.

[15] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, Kilian Q Weinberger: Densely connected convolutional networks. In: Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4700-4708.

[16] Q. Guan, Y. Huang, Multi-label chest X-ray image classification via category-wise residual attention learning, Pattern Recog Lett. 130 (2020) 259–266.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, Illia Polosukhin: Attention is all you need. In: Neural Information Processing Systems (NIPS), 2017, pp. 5998-6008.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv: 2010.11929.

[19] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, Ling Shao, Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, 2021, arXiv preprint arXiv: 2102.12122.

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo, Swin transformer: Hierarchical vision transformer using shifted windows, 2021, arXiv preprint arXiv: 2103.14030.

[21] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Xiyang Dai, Bin Xiao, Lu Yuan, Jianfeng Gao, Focal Self-attention for Local-Global Interactions in Vision Transformers, 2021, arXiv preprint arXiv: 2107.00641.

[22] Ido Freeman, Lutz Roese-Koerner, Anton Kummert: Effnet: An efficient structure for convolutional neural networks. In: 2018 25th ieee international conference on image processing (icip), 2018, pp. 6-10.

[23] I. Sluimer, A. Schilham, M. Prokop, B. Van Ginneken, Computer analysis of computed tomography scans of the lung: a survey, IEEE Trans Med Imaging. 25 (4) (2006) 385–405.

[24] F. Li, S. Sone, H. Abe, H. MacMahon, Samuel G Armato, Kunio Doi, Lung cancers missed at low-dose helical CT screening in a general population: comparison of clinical, histopathologic, and imaging findings, Radiology 225 (3) (2002) 673–683.

[25] S.W. Davies, Clinical presentation and diagnosis of coronary artery disease: stable angina, Br Med Bull. 59 (1) (2001) 17–27.

[26] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, 2017, arXiv preprint arXiv: 1711.05225.

[27] H. Wang, H. Jia, L.u. Le, Y. Xia, Thorax-net: an attention regularized deep neural network for classification of thoracic diseases on chest radiography, IEEE J. Biomed. Health Inform. 24 (2) (2019) 475–485.

[28] Jaehyup Jeong, Bosoung Jeoun, Yeonju Park, Bohyung Han: An Optimized Ensemble Framework for Multi-Label Classification on Long-Tailed Chest X-ray Data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2739-2746.

[29] Y. Jin, L.u. Huijuan, W. Zhu, W. Huo, Deep learning based classification of multi-label chest X-ray images via dual-weighted metric loss, Computers in Biology Medicine. 157 (2023) 106683.

[30] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, Ruslan Salakhutdinov, Transformer-xl: Attentive language models beyond a fixed-length context, 2019, arXiv preprint arXiv:1901.02860.

[31] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.

[32] K. Lagler, M. Schindelegger, J. Böhm, H. Krásná, T. Nilsson, GPT2: Empirical slant delay model for radio space geodetic techniques, Geophys Res Lett. 40 (6) (2013) 1069–1073.

[33] A. Jamali, Swalpa Kumar Roy, Avik Bhattacharya, Pedram Ghamisi, Local window attention transformer for polarimetric SAR image classification, IEEE Geoscience, Remote Sensing Letters. 20 (2023) 1–5.

[34] C. Zhao, B. Qin, S. Feng, W. Zhu, W. Sun, W. Li, X. Jia, Hyperspectral image classification with multi-attention transformer and adaptive superpixel segmentation-based active learning, IEEE Trans. Image Process. 32 (2023) 3606–3621.

[35] O. Russakovsky, J. Deng, S.u. Hao, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.

[36] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, Lihi Zelnik-Manor: Asymmetric loss for multi-label classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 82-91.

[37] Ilya Loshchilov, Frank Hutter, Sgdr: Stochastic gradient descent with warm restarts, 2016, arXiv preprint arXiv:1608.03983.

[38] Ilya Loshchilov, Frank Hutter, Decoupled weight decay regularization, 2017, arXiv preprint arXiv: 1711.05101.

[39] B. Chen, J. Li, X. Guo, L.u. Guangming, Control, DualCheXNet: dual asymmetric feature learning for thoracic disease classification in chest X-rays, Biomedical Signal Processing Control. 53 (2019) 101554.

[40] B. Chen, J. Li, L.u. Guangming, D. Zhang, Lesion location attention guided network for multi-label thoracic disease classification in chest X-rays, IEEE Journal of Biomedical Health Informatics. 24 (7) (2019) 2016–2027.

[41] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, Ling Shao, PVT v2: Improved baselines with Pyramid Vision Transformer, Computational Visual Media. (2022) 1-10.

[42] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, M. Müller, pROC: an open-source package for R and S+ to analyze and compare ROC curves, BMC Bioinf. 12 (1) (2011) 1–8.

[43] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618-626.

[44] L. Van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008) 2579–2605.