Revisiting Computer-Aided Tuberculosis Diagnosis

Yun Liu, Yu-Huan Wu, Shi-Chen Zhang, Li Liu, Min Wu, and Ming-Ming Cheng

Abstract—Tuberculosis (TB) is a major global health threat, causing millions of deaths annually. Although early diagnosis and treatment can greatly improve the chances of survival, it remains a major challenge, especially in developing countries. Recently, computer-aided tuberculosis diagnosis (CTD) using deep learning has shown promise, but progress is hindered by limited training data. To address this, we establish a large-scale dataset, namely the Tuberculosis X-ray (TBX11K) dataset, which contains 11,200 chest X-ray (CXR) images with corresponding bounding box annotations for TB areas. This dataset enables the training of sophisticated detectors for high-quality CTD. Furthermore, we propose a strong baseline, SymFormer, for simultaneous CXR image classification and TB infection area detection. SymFormer incorporates Symmetric Search Attention (SymAttention) to tackle the *bilateral symmetry property* of CXR images for learning discriminative features. Since CXR images may not strictly adhere to the bilateral symmetry property, we also propose Symmetric Positional Encoding (SPE) to facilitate SymAttention through feature recalibration. To promote future research on CTD, we build a benchmark by introducing evaluation metrics, evaluating baseline models reformed from existing detectors, and running an online challenge. Experiments show that SymFormer achieves state-of-the-art performance on the TBX11K dataset. The data, code, and models will be released at https://github.com/yun-liu/Tuberculosis.

Index Terms—Tuberculosis, tuberculosis diagnosis, tuberculosis detection, symmetric search attention, symmetric positional encoding

✦

1 INTRODUCTION

TUBERCULOSIS (**TB**), a pervasive infectious disease, has persistently ranked as the second leading cause of morbidity and mortality, typically following HIV, over the centuries [2], [3]. Despite the global COVID-19 outbreak in 2020, TB continues to afflict 10 million individuals and accounts for the death of 1.4 million people annually [4], rendering it the second most lethal infectious disease after COVID-19. Principally targeting the respiratory system, TB is caused by Mycobacterium tuberculosis and propagates through sneezing, severe coughing, or other means of disseminating infectious bacteria. Hence, TB typically occurs in the lungs through the respiratory tract. The vulnerability of immunocompromised individuals, including those with HIV and malnourished persons in developing countries, has exacerbated this issue.

The mortality rate among TB patients remains exceedingly high in the absence of appropriate treatment. Nevertheless, early diagnosis of TB can significantly increase the recovery rate with the administration of corresponding antibiotics [5]–[7]. As TB propagates rapidly, early diagnosis also plays a crucial role in controlling the spread of infection [6]. The rise of multidrug-resistant TB underscores the urgent need for timely and accurate diagnostic methods to monitor the progress of clinical treatment [8]. However, TB diagnosis continues to pose a significant challenge [5]–[7], [9]–[12]. Specifically, the *gold standard* for TB diagnosis entails the microscopic examination of sputum samples and bacterial cultures to identify Mycobacterium tuberculosis [11], [12]. To ensure the safety of the examination process, a biosafety level-3 (BSL-3) laboratory is required for culturing Mycobacterium tuberculosis. This procedure can typically take several months [5], [11], [12]. Compounding the issue, many hospitals in developing countries and resource-constrained communities *lack the necessary infrastructure* to establish BSL-3 facilities.

On the other hand, X-ray imaging is the most prevalent and data-intensive screening method in current medical image examinations. Chest X-ray (**CXR**) can swiftly detect lung abnormalities caused by pulmonary TB, making it a widely-used tool for TB screening. The World Health Organization also recommends CXR as the initial step in TB screening [13]. Early diagnosis through CXR significantly aids in early TB detection, treatment, and prevention of the disease's spread [5], [10], [13]–[15]. However, even experienced radiologists may fail to identify TB infections in CXR images, as the human eye struggles to discern TB areas in CXR images due to its limited sensitivity to many details. Our human study reveals that experienced radiologists from top hospitals achieve *an accuracy of only 68.7%* when compared with the gold standard.

Thanks to the remarkable representation learning capabilities, deep learning has outperformed humans in various domains such as face recognition [16], image classification

• *This work was mostly done when Y. Liu was with VCIP, Nankai University, Tianjin, China. He is now with the Institute for Infocomm Research (I2R), A*STAR, Singapore. (E-mail: vagrantlyun@gmail.com)*
• *Y.-H. Wu is with the Institute of High Performance Computing (IHPC), A*STAR, Singapore. (E-mail: wu_yuhuan@ihpc.a-star.edu.sg)*
• *S.-C. Zhang and M.-M. Cheng are with VCIP, Nankai University, Tianjin, China. (E-mail: zhangshichen@mail.nankai.edu.cn and cmm@nankai.edu.cn)*
• *L. Liu is with the College of Electronic Science and Technology, National University of Defense Technology, Changsha, Hunan, China. (E-mail: liuli_nudt@nudt.edu.cn)*
• *M. Wu is with the Institute for Infocomm Research (I2R), A*STAR, Singapore. (E-mail: wumin@i2r.a-star.edu.sg)*
• *Corresponding author: Li Liu. (E-mail: liuli_nudt@nudt.edu.cn)*
• *A preliminary version of this work has been published on CVPR (oral) [1].*

TABLE 1
Summary of publicly available TB datasets. The size of our dataset is about $17\times$ larger than that of the previous largest dataset. Besides, our dataset annotates TB infection areas with bounding boxes, instead of only image-level labels.

Datasets	Pub. Year	#Classes	Annotations	#Samples
MC [25]	2014	2	Image-level	138
Shenzhen [25]	2014	2	Image-level	662
DA [6]	2014	2	Image-level	156
DB [6]	2014	2	Image-level	150
TBX11K (Ours)	-	4	Bounding box	11,200

[17], object detection [18], [19], edge detection [20], [21], and medical image analysis [22]–[24]. It is reasonable to anticipate the application of deep learning's robust potential to TB diagnosis using CXR. Deep learning can automatically localize the precise TB infection site 24 hours a day, never getting tired like humans. However, deep learning relies on extensive training data, which cannot be provided by existing TB datasets, as shown in Table 1. Since it is challenging to collect large-scale TB CXR data due to the high cost and privacy considerations, existing TB datasets have only a few hundred CXR images. The scarcity of publicly available CXR data has *hindered* the successful application of deep learning in improving computer-aided tuberculosis diagnosis (**CTD**) performance.

In order to deploy the CTD system to assist TB patients worldwide, it is first necessary to address the issue of insufficient data. In this paper, we contribute a large-scale **Tuberculosis X-ray (TBX11K)** dataset to the community through long-term collaboration with major hospitals. This new TBX11K dataset surpasses previous CTD datasets in several aspects: i) Unlike previous public datasets [6], [25] containing only tens or hundreds of CXR images, TBX11K consists of 11,200 CXR images, approximately 17 times larger than the existing largest dataset, *i.e.*, the Shenzhen dataset [25], making it feasible to train deep networks; ii) In contrast to image-level annotations in previous datasets, TBX11K employs bounding box annotations for TB infection areas, allowing future CTD methods to recognize TB manifestations and detect TB regions for assisting radiologists in definitive diagnoses; iii) TBX11K comprises four categories: healthy, sick but non-TB, active TB, and latent TB, as opposed to binary classification in previous datasets (*i.e.*, TB or non-TB), enabling future CTD systems to adapt to more complex real-world scenarios and provide people with more detailed disease analyses. Each CXR image in the TBX11K dataset is tested using the gold standard (*i.e.*, diagnostic microbiology) of TB diagnosis and annotated by experienced radiologists from major hospitals. The TBX11K dataset has been de-identified by data providers and exempted by relevant institutions, allowing it to be publicly available to promote future CTD research.

Based on our TBX11K dataset, we propose a simple yet effective framework for CTD, termed as **SymFormer**. Inspired by the inherent *bilateral symmetry property* observed in CXR images, SymFormer leverages this property to enhance the interpretation of CXR images. The bilateral symmetry property denotes the similarity or identical appearance of the left and right sides of the chest, indicating a symmetric

pattern. This property proves valuable in improving the interpretation of CXR images. For instance, if there is a mass or consolidation present on one side of the chest but not the other, it could indicate a problem in that area. To tackle this property, SymFormer incorporates the novel **Symmetric Search Attention (SymAttention)** for learning discriminative features from CXR images. Since CXR images may not strictly be bilaterally symmetric, we also propose the **Symmetric Positional Encoding (SPE)** to facilitate SymAttention through feature recalibration. SymFormer conducts simultaneous CXR image classification and TB infection area detection by adding a classification head onto the TB infection area detector with a two-stage training diagram.

To promote future research on CTD, we establish a benchmark on our TBX11K dataset. Specifically, we adapt the evaluation metrics for image classification and object detection to CTD, which would standardize the evaluation of CTD. We also launch an online challenge using the test data of TBX11K by keeping the ground truth of the test data private, which would make future comparisons on CTD fair. Besides, we construct several strong baseline models for CTD by reforming existing popular object detectors. Extensive comparisons demonstrate the superiority of SymFormer over these baselines.

Compared with the preliminary conference version [1], we make plentiful extensions by proposing a novel SymFormer framework for CTD and validating its effectiveness with extensive experiments. In summary, the contributions of this paper are three-fold:

- We establish a large-scale CTD dataset, TBX11K, which is much larger, better annotated, and more realistic than previous TB datasets, enabling the training of deep neural networks for simultaneous multiclass CXR image classification and TB infection area detection rather than only binary CXR classification in previous TB datasets.
- We propose a simple yet effective framework for CTD, namely SymFormer, consisting of the novel Symmetric Search Attention (SymAttention) and Symmetric Positional Encoding (SPE) to leverage the *bilateral symmetry property* of CXR images for significantly improving CTD over baseline models.
- We build a CTD benchmark on our TBX11K dataset by introducing the evaluation metrics, evaluating several baselines reformed from existing object detectors, and running an online challenge, which is expected to set a good start for future research.

2 RELATED WORK

In this section, we first revisit previous TB datasets, followed by a review of the existing research on CTD. Since our proposed CTD method SymFormer uses self-attention of vision transformers, we also discuss the recent progress of vision transformers in medical imaging.

2.1 Tuberculosis Datasets

Since TB data are very private and it is difficult to diagnose TB with the golden standard, the publicly available TB datasets are very limited. We provide a summary for the

publicly available TB datasets in Table 1. Jaeger *et al.* [25] established two CXR datasets for TB diagnosis. The Montgomery County chest X-ray set (MC) [25] was collected through cooperation with the Department of Health and Human Services, Montgomery County, Maryland, USA. MC dataset consists of 138 CXR images, 80 of which are healthy cases and 58 are cases with manifestations of TB. The Shenzhen chest X-ray set (Shenzhen) [25] was collected through cooperation with Shenzhen No. 3 People's Hospital, Guangdong Medical College, Shenzhen, China. The Shenzhen dataset is composed of 326 norm cases and 336 cases with manifestations of TB, leading to 662 CXR images in total. Chauhan *et al.* [6] built two datasets, namely DA and DB, which were obtained from two different X-ray machines at the National Institute of Tuberculosis and Respiratory Diseases, New Delhi. DA is composed of the training set (52 TB and 52 non-TB CXR images) and the independent test set (26 TB and 26 non-TB CXR images). DB contains 100 training CXR images (50 TB and 50 non-TB) and 50 test CXR images (25 TB and 25 non-TB). Note that all these four datasets are annotated with image-level labels for binary CXR image classification.

These datasets are too small to train deep neural networks, so recent research on CTD has been hindered although deep learning has achieved numerous success stories in the computer vision community. On the other hand, the existing datasets only have image-level annotations, and thus we cannot train TB detectors with previous data. To help radiologists make accurate diagnoses, we are expected to detect the TB infection areas, not only an image-level classification. Therefore, the lack of TB data has prevented deep learning from bringing success to practical CTD systems that have the potential to save millions of TB patients every year. In this paper, we build a large-scale dataset with bounding box annotations for training deep neural networks for simultaneous CXR image classification and TB infection area detection. The presentation of this new dataset is expected to benefit future research on CTD and promote more practical CTD systems.

2.2 Computer-aided Tuberculosis Diagnosis

Owing to the lack of data, traditional CTD methods cannot train deep neural networks. Thus, traditional methods mainly use hand-crafted features and train binary classifiers for CXR image classification. Jaeger *et al.* [5] first segmented the lung region using a graph cut segmentation method [26]. Then, they extracted hand-crafted texture and shape features from the lung region. Finally, they apply a binary classifier, *i.e.*, support vector machine (SVM), to classify the CXR image as normal or abnormal. Candemir *et al.* [10] adopted image retrieval-based patient-specific adaptive lung models to a nonrigid registration-driven robust lung segmentation method, which would be helpful for traditional lung feature extraction [5]. Chauhan *et al.* [6] implemented a MATLAB toolbox, TB-Xpredict, which adopted Gist [27] and PHOG [28] features for the discrimination between TB and non-TB CXR images without requiring segmentation [29], [30]. Karargyris *et al.* [31] extracted shape features to describe the overall geometrical characteristics of lungs and texture features to represent image characteristics.

Instead of using hand-crafted features, Lopes *et al.* [9] adopted the frozen convolutional neural networks pretrained on ImageNet [32] as the feature extractors for CXR images. Then, they train SVM to classify the extracted deep features. Hwang *et al.* [7] trained an AlexNet [33] for binary classification (TB and non-TB) using a private dataset. Other private datasets are also used in [34] for image classification networks. However, our proposed large-scale dataset, *i.e.*, TBX11K, has been made publicly available to promote research in this field. With our new dataset, we propose a transformer-based CTD method, SymFormer, for simultaneous CXR image classification and TB infection area detection, which serves as a strong baseline for future research on CTD by achieving state-of-the-art performance.

2.3 Vision Transformers in Medical Imaging

Transformer [35] is initially introduced in natural language processing (NLP), and it has a good ability to capture long-range dependencies. Pioneering works on adapting transformers to vision tasks, such as ViT [36], DeiT [37], and P2T [38], showed that transformer networks can surpass the widely-used convolutional neural networks. Therefore, vision transformers attract increasing attention from the computer vision community, including medical imaging. Various efforts have been made to incorporate vision transformers into medical image segmentation [39]–[43] and medical image classification [44]–[48]. However, the adoption of transformer-based techniques for medical image detection lags behind that of segmentation and classification.

Most studies utilizing vision transformers for medical image detection are primarily built on the detection transformer (DETR) framework [49]. The pioneering work in this field is COTR [50], comprising a convolutional neural network for feature extraction, hybrid convolution-in-transformer layers for feature encoding, transformer decoder layers for object querying, and a feed-forward network for polyp detection. Mathai *et al.* [51] employed DETR [49] to detect lymph nodes in T2 MRI scans, which can be used to evaluate lymphoproliferative diseases. Li *et al.* [52] proposed a Slice Attention Transformer (SATr) block to model the long-range dependency among different computed tomography (CT) slices, which can be plugged into convolution-based models for universal lesion detection. Please refer to recent survey papers [53]–[55] for a more comprehensive review of vision transformers in medical imaging. In this paper, we propose SymFormer for CTD using CXR images. SymFormer conducts simultaneous CXR image classification and TB infection area detection. It leverages SymAttention to tackle the *bilateral symmetry property* of CXR images, which is further promoted by SPE. With SymAttention and SPE, SymFormer exhibits much better performance than recent popular object detector baselines, suggesting its superiority in CTD.

3 TBX11K DATASET

Deep neural networks are highly dependent on large amounts of training data, while existing public TB datasets are not large-scale as shown in Table 1. To address this issue, we establish a comprehensive and large-scale dataset called
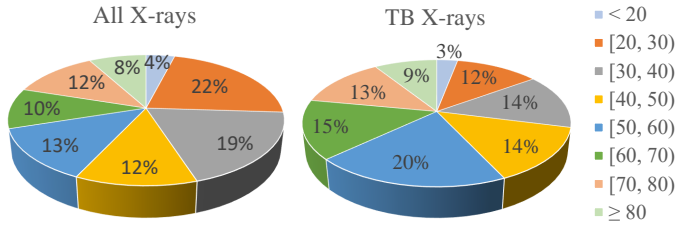
Fig. 1. Age distributions for the entire TBX11K dataset and specifically for TBX11K TB X-rays.
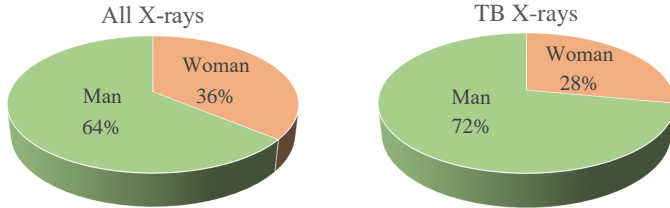

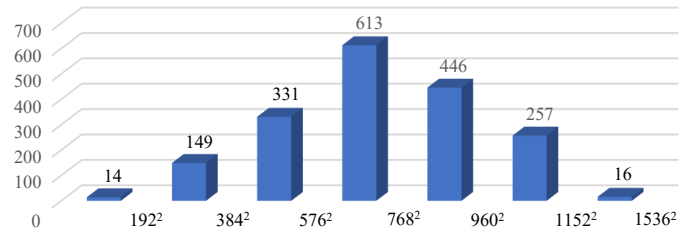Fig. 2. Gender distributions for the entire TBX11K dataset and specifically for TBX11K TB X-rays.

Fig. 3. **Distribution of the sizes of TB bounding boxes in the TBX11K dataset.** Each bin represents a specific range of bounding box areas. The left and right values of each bin correspond to its area range, and the height of the bin represents the number of TB bounding boxes within that range. It should be noted that the CXR images in TBX11K have a resolution of about 3000×3000.

TBX11K, which enables the training of deep networks for CTD. In this section, we first describe how we collect and annotate the CXR data in §3.1. Next, we present the results of a human study conducted by experienced radiologists in §3.2. Finally, we discuss potential research topics that can be explored using our TBX11K dataset in §3.3.

3.1 Data Collection and Annotation

To collect and annotate the data, we adhere to four primary steps: i) establishing a taxonomy, ii) collecting CXR data, iii) professional data annotation, and iv) dataset splitting. We will introduce each of these steps in detail below.

3.1.1 Taxonomy Establishment

The current TB datasets only consist of two categories: TB and non-TB, where non-TB refers to healthy cases. However, in practice, abnormalities in CXR images that indicate TB, atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, *etc.*, have similar abnormal patterns such as blurry and irregular lesions, which differ significantly from healthy CXR that have almost clear patterns. Therefore, relying solely on healthy CXR as the negative category leads to biases that can cause large false positives in the model's prediction for clinical scenarios where there are many sick but non-TB patients. To address this issue and promote the practical application of CTD, we propose a new category, sick but non-TB, in our dataset. Furthermore, differentiating between active TB and latent TB is crucial in providing patients with proper treatment. Active TB results from Mycobacterium TB infection or reactivation of latent TB, while individuals with latent TB are neither sick nor contagious. Therefore, we have divided TB into two categories of active TB and latent TB in our dataset. In light of the above analysis, the proposed TBX11K dataset includes four categories: healthy, sick but non-TB, active TB, and latent TB.

3.1.2 Data Collection

The collection of TB CXR data presents two main challenges: i) The high privacy of CXR data, particularly TB CXR data,

making it almost impossible for individuals to access the raw data without risking breaking the law; ii) The scarcity of definitively tested TB CXR images, due to the complex and lengthy process of examining Mycobacterium TB using the golden standard [11], [12], despite the millions of TB patients worldwide. To address these challenges, we collaborate with top hospitals in China to gather the CXR data. Our resulting TBX11K dataset comprises 11,200 CXR images, including 5,000 healthy cases, 5,000 sick but non-TB cases, and 1,200 TB cases. Each CXR image corresponds to a unique individual. Of the 1,200 TB CXR images, 924 are active TB cases, 212 are latent TB cases, 54 contain both active and latent TB, and 10 are uncertain cases whose TB types cannot currently be recognized. We include 5,000 sick but non-TB cases to cover a broad range of radiograph diseases that can appear in clinical scenarios. The data providers have de-identified the data, and relevant government institutions have exempted the dataset, making it publicly available legally.

All CXR images are in a resolution of approximately 3000×3000. Each CXR image is accompanied by corresponding age and gender information, providing comprehensive clinical clues for the diagnosis of TB. The age distribution and gender distribution are presented in Fig. 1 and Fig. 2, respectively. From Fig. 1, it is evident that 60% of the TB patients fall within the age range of 20 to 60 years. Only a small percentage of young individuals (aged less than 20) are infected with TB, specifically 3% of TB patients. This finding is consistent with recent medical research [56]–[58]. Fig. 2 illustrates that the majority of TB patients are male, aligning with clinical observations that TB is more prevalent in men than women globally [59], [60]. We have annotated TB infection areas using bounding boxes (introduced in §3.1.3), and the distribution of the sizes of these TB bounding boxes is depicted in Fig. 3. As evident from the figure, TB infection areas exhibit a wide range of sizes, manifesting varying degrees of TB severity. Taken together with the previous analyses on taxonomy, category distribution, age distribution, gender distribution, and TB infection size distribution, we can come to the conclusion that the new TBX11K dataset is *both representative of the general population and consistent with real-world clinical scenarios.*

3.1.3 Professional Data Annotation

Our dataset comprises CXR images that have undergone rigorous testing using the golden standard, which provides

TABLE 2
Split for the TBX11K dataset. "Active & Latent TB" refers to CXR images with both active and latent TB; "Active TB" refers to CXR images with only active TB; "Latent TB" refers to CXR images with only latent TB; "Uncertain TB" refers to TB CXR images where the type of TB infection cannot be recognized using current medical conditions.

	Classes	Train	Val	Test	Total
Non-TB	Healthy	3,000	800	1,200	5,000
	Sick & Non-TB	3,000	800	1,200	5,000
TB	Active TB	473	157	294	924
	Latent TB	104	36	72	212
	Active & Latent TB	23	7	24	54
	Uncertain TB	0	0	10	10
Total		6,600	1,800	2,800	11,200

image-level labels. However, while this approach enables us to categorize a CXR image as indicative of TB if the sputum of the corresponding patient shows manifestations of the disease, it does not reveal the specific location or extent of the TB in the CXR image. The ability to detect these TB infection areas is crucial to enable radiologists to make informed decisions. Currently, relying solely on image-level predictions makes it difficult for the human eye to identify TB infection areas, as evidenced by the low accuracy of radiologists during clinical examinations (see §3.2). By simultaneously providing image classification and TB localization results, CTD systems have the potential to enhance the accuracy and efficiency of radiologists in making informed decisions.

To achieve our goal, our TBX11K dataset includes bounding box annotations for TB infection areas in CXR images. To the best of our knowledge, this is the first dataset designed for TB infection area detection. These annotations are carried out by experienced radiologists from top hospitals. Specifically, each TB CXR image in the dataset is first labeled by a radiologist with 5-10 years of experience in TB diagnosis. Subsequently, another radiologist with over 10 years of experience in TB diagnosis reviews the box annotations. The radiologists do not just label bounding boxes for TB areas but also identify the type of TB (active or latent) for each box. To ensure consistency, the labeled TB types are double-checked against the image-level labels produced by the golden standard. In the event of a mismatch, the CXR image is placed in the unlabeled data for re-annotation, and the annotators do not know which CXR image was previously labeled incorrectly. If a CXR image is labeled incorrectly twice, we inform the annotators of the gold standard for that CXR image and request that they discuss how to re-annotate it. This double-checked process ensures that the annotated bounding boxes are highly reliable for TB infection area detection. Additionally, non-TB CXR images are only labeled with image-level labels produced by the golden standard. Examples of the TBX11K dataset are shown in Fig. 6, and the distribution of TB bounding box sizes is displayed in Fig. 3, indicating that most TB bounding boxes are in the range of $(384^2, 960^2]$.

3.1.4 Dataset Splitting

We have partitioned the data into three subsets: training, validation, and test, following the split detailed in Table 2.

The ground truths for both the training and validation sets have been made public, whereas the ground truths for the test set remain confidential. This is because we have launched an online challenge using the test data on our website. To ensure a more representative dataset, we have considered four distinct TB cases: i) CXR images with active TB only, ii) CXR images with latent TB only, iii) CXR images with both active and latent TB, and iv) CXR images with uncertain TB type that cannot be recognized under current medical conditions. For each TB case, we have maintained a ratio of $3 : 1 : 2$ for the number of TB CXR images in the training, validation, and test sets. It is worth noting that the uncertain TB CXR images have been assigned to the test set, enabling researchers to evaluate class-agnostic TB detection using these 10 uncertain CXR images. We recommend that researchers train their models on the training set, tune hyper-parameters on the validation set, and report the model's performance on the test set after retraining using the union of the training and validation sets. This approach follows scientific experiment settings and is expected to yield reliable results.

3.2 Human Study by Radiologists

The human study involving radiologists is a critical component in understanding the role of CTD in clinical settings. We begin by randomly selecting 400 CXR images from the test set of the new TBX11K dataset, which includes 140 healthy CXR images, 140 sick but non-TB CXR images, and 120 CXR images with TB. Of the 120 CXR images with TB, 63 show active TB, 41 show latent TB, 15 show both active and latent TB, and 1 shows uncertain TB. Next, we invite an experienced radiologist from a major hospital with over 10 years of work experience to label the CXR images according to four image-level categories: healthy, sick but non-TB, active TB, and latent TB. If a CXR image displays both active and latent TB manifestations, the radiologist assigns both labels. It is important to note that this radiologist is different from those who labeled the original dataset.

The radiologist achieves an accuracy of only 68.7% when compared to the ground truth produced by the golden standard. If we ignore the differentiation between active and latent TB, the accuracy improves to 84.8%, but distinguishing between the types of TB is crucial for effective clinical treatment. This low performance highlights one of the major challenges in TB diagnosis, treatment, and prevention. Unlike natural color images, CXR images are grayscale and often have fuzzy and blurry patterns, making accurate recognition challenging. Unfortunately, diagnosing TB with the golden standard can take several months in a BSL-3 laboratory [11], [12], which is not feasible in many parts of the world. The challenge in TB diagnosis leads to TB becoming the second most common infectious disease worldwide after HIV. However, we will show in our upcoming study that deep-learning-based CTD models trained on the proposed TBX11K dataset can significantly outperform even experienced radiologists, offering hope for improved TB diagnosis and treatment.

3.3 Potential Research Topics

Moving forward, we discuss some potential research topics related to CTD using our newly developed TBX11K dataset.

Simultaneous classification and detection. Our TBX11K dataset opens up new possibilities for conducting research on CTD, including CXR image classification and TB infection area detection. Our test set includes a broad range of health and non-TB sick data, enabling the simulation of clinical data distribution for evaluating CTD systems. We believe that the development of simultaneous CXR image classification and TB infection area detection systems would be a challenging and fascinating research topic, with potential applications for assisting radiologists in TB diagnosis. Deploying such systems could ultimately improve the accuracy and efficiency of TB diagnosis and treatment.

Imbalanced data distribution. In addition to the challenge of simultaneous detection and image classification, our TBX11K dataset also presents an imbalanced data distribution across different categories. However, we believe that this data imbalance is reflective of real-world clinical scenarios. When patients undergo chest examinations, they may be experiencing discomfort or illness, increasing the likelihood of getting sick, and our dataset captures this reality with only 44.6% of takers being healthy. TB is just one of many possible chest diseases, and our dataset reflects this reality with only 10.7% of takers being infected with TB, while 44.6% are sick but non-TB. Latent TB can result from two scenarios: exposure to active TB and conversion from active TB after treatment. Most cases of latent TB are caused by exposure to active TB. However, individuals with latent TB are not sick or contagious and are unlikely to seek medical attention, resulting in a higher number of active TB cases in our dataset than latent TB cases. This data imbalance presents a challenge for future CTD methods, which must be designed to overcome this problem in practice. For example, methods for training models on the imbalanced TBX11K training set will need to be developed to improve the accuracy of TB diagnosis.

Incremental learning with private data. Incremental learning is a machine learning technique that involves updating a model's parameters with new data as it becomes available, without requiring the model to be retrained from scratch. Given the high privacy concerns surrounding TB CXR data, researchers may possess private data that cannot be released. In such cases, it may be beneficial to use a model pre-trained on the TBX11K dataset as the base model. Researchers can then leverage incremental learning to fine-tune the pre-trained model using their private data, thereby enhancing the model's capacity for accurate CTD. Hence, investigating the potential of incremental learning for CTD using the newly developed TBX11K dataset would also be a crucial research direction.

4 OUR SYMFORMER FRAMEWORK

In this section, we first present an overview of our Sym-Former framework in §4.1. Then, we describe our **Symmetric Abnormality Search (SAS)** method in §4.2. SAS consists of two components: **Symmetric Positional Encoding (SPE)** in §4.2.1, and **Symmetric Search Attention (SymAttention)** in §4.2.2. Next, we introduce the TB diagnosis heads for Sym-Former in §4.3. Finally, we present the two-stage training
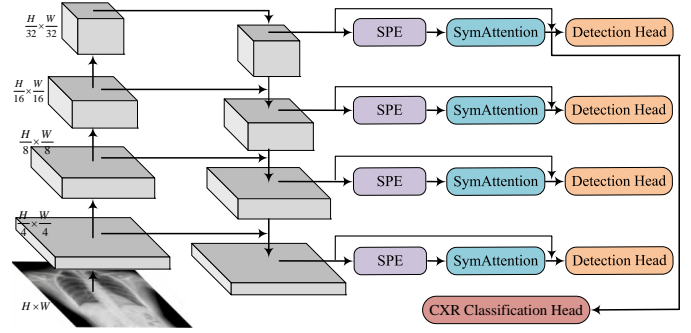

Fig. 4. **Illustration of the proposed SymFormer framework.** FPN [61] is applied to generate the feature pyramid.

diagram for simultaneous CXR image classification and TB infection area detection in §4.4.

4.1 Overview

We illustrate the overall pipeline of SymFormer in Fig. 4. SymFormer comprises three parts: feature extraction, symmetric abnormity search, and TB diagnosis heads. We will elaborate on each part below.

Feature extraction. For the sake of convenience, we take ResNets [62] as an example backbone network for feature extraction due to its generality acknowledged by the community. When given a CXR image as input, the backbone network outputs features in four stages, which are scaled down by factors of 1/4, 1/8, 1/16, and 1/32, respectively, in comparison to the input size. As the sizes and shapes of TB infection areas vary widely, it is crucial to capture multi-scale features from the backbone network. In order to achieve this, a feature pyramid network (FPN) [61] is applied upon the backbone network, which generates a feature pyramid, *i.e.*, feature maps at different scales. We denote the feature pyramid as $\mathbf{F} = \{\mathbf{F}_1, \mathbf{F}_2, \mathbf{F}_3, \mathbf{F}_4\}$ *w.r.t.* $\mathbf{F}_i \in \mathbb{R}^{C \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}}$ ($i \in \{1, 2, 3, 4\}$), in which C is the feature dimension and H and W are the height and width of the input CXR image, respectively. The feature pyramid is effective at enabling TB infection detection across different feature levels.

Symmetric abnormity search. The SAS module serves to enhance the extracted feature pyramid $\mathbf{F}$. To achieve this, an SAS module is incorporated after each side output of FPN [61] to process each feature map $\mathbf{F}_i$ in the feature pyramid $\mathbf{F}$. The enhanced feature pyramid is expressed as $\hat{\mathbf{F}} = \{\hat{\mathbf{F}}_1, \hat{\mathbf{F}}_2, \hat{\mathbf{F}}_3, \hat{\mathbf{F}}_4\}$ *w.r.t.* $\hat{\mathbf{F}}_i \in \mathbb{R}^{C \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}}$ ($i \in \{1, 2, 3, 4\}$). The SAS modules at various side outputs share the same weights to reduce the number of network parameters. According to the *bilateral symmetry property*, the bilaterally symmetric regions in a normal CXR image should look similar or identical. The SAS module leverages this insight by searching for symmetric positions in each position of the feature map to determine if it is normal. The SAS module consists of three components: SPE, SymAttention, and a feed-forward network. While the CXR image may not be strictly symmetric, the SPE is designed to recalibrate the fea-

tures, which then benefits the SymAttention for symmetric-search-based feature enhancement.

TB diagnosis heads. We connect two types of TB diagnosis heads to the feature pyramid $\hat{\mathbf{F}}$, which is enhanced by the SAS module, for performing TB infection area detection and CXR image classification, respectively. Each feature map in the feature pyramid $\hat{\mathbf{F}}$ is fed into the detection head, and each detected bounding box is expected to cover a TB infection area. However, there is a risk of introducing false positives for non-TB CXR images during TB infection area detection, which leads to unnecessary costs for radiologists to check these false positives for clinical diagnosis. To address this issue, we feed the feature map $\hat{\mathbf{F}}_4$ at the top level of the enhanced feature pyramid into a classification head to determine whether a CXR image contains TB or not. If a CXR image is classified as TB, radiologists can further examine the detected TB infection areas for a more accurate and detailed clinical diagnosis. If a CXR image is classified as non-TB, the detected areas need not be checked further by radiologists.

4.2 Symmetric Abnormality Search

Bilateral symmetry is a property of CXR images where the structures on the left and right sides of the chest appear similar or identical. In other words, if a line is drawn down the center of the CXR image, the structures on either side of the line should be approximately the same size and shape. This property plays a crucial role in the interpretation of CXR images since it enables radiologists and clinicians to identify asymmetries or abnormalities in the lung fields. For example, the presence of a mass or consolidation on one side of the lung but not the other could indicate a problem in that area. However, it is worth noting that perfect bilateral symmetry is not always present in normal CXR images, depending on the patient's pose and position relative to the X-ray machine when the CXR image is taken.

Our proposed method, SAS, leverages the bilateral symmetry property to enhance the feature representations of CXR images. As mentioned above, the lungs in CXR images may not be strictly symmetric. To account for this, SAS first incorporates SPE for feature recalibration. This recalibrated feature map is then used by SymAttention to search the symmetric adjacent area of each spatial position in the feature map, where the symmetric adjacent area refers to the adjacent area of the bilaterally symmetric position for a given position. SymAttention aggregates features in the symmetric adjacent area in an adaptive way through attention. The adjacent area is also determined in a learning way. By forcing each spatial position to look at the symmetric adjacent area, as suggested by the bilateral symmetry property, we can learn discriminative features for the CXR image for CTD.

4.2.1 Symmetric Positional Encoding

To incorporate positional information into self-attention computations for a feature map, we must add positional encoding to the feature map. There are two types of positional encoding: absolute positional encoding and relative positional encoding [35], [36]. Our method, called SPE, is

based on absolute positional encoding, and our experiments indicate that relative positional encoding is inferior to our SPE, as shown in §6.3. The widely-used absolute positional encoding [35], [36] employs sine and cosine functions of different frequencies:

$$
\begin{aligned}
\mathbf{P}[pos, 2j] &= \sin(pos/10000^{\frac{2j}{C}}), \\
\mathbf{P}[pos, 2j+1] &= \cos(pos/10000^{\frac{2j}{C}}),
\end{aligned}
\tag{1}
$$

where pos denotes the spatial position and j indexes the feature dimension. For each input feature map $\mathbf{F}_i$ from the feature pyramid $\mathbf{F}$, we use Eq. 1 to calculate the corresponding positional encoding $\mathbf{P}_i$. $\mathbf{P}_i$ has the same shape as $\mathbf{F}_i$ so that $\mathbf{P}_i$ and $\mathbf{F}_i$ can be summed.

As mentioned earlier, CXR images may not strictly adhere to the bilateral symmetry property as they can have slight rotations and translations. The proposed SPE is designed to tackle this issue by feature recalibration. SPE first splits the positional encoding $\mathbf{P}_i$ into two sides, i.e., $\mathbf{P}_i^{left}$ and $\mathbf{P}_i^{right}$, by drawing a line down the center of $\mathbf{P}_i$. Then, we transfer $\mathbf{P}_i^{right}$ to the left side using spatial transformer networks (STN) [63] and horizontal flipping. Finally, we concatenate the transformed left-side positional encoding and $\mathbf{P}_i^{right}$ along the x dimension to form the output $\mathbf{P}_i^{sym}$. This process can be formulated as follows:

$$
\begin{aligned}
\mathbf{T}_i &= \text{STN}(\mathbf{F}_i; \Theta) \\
\mathbf{P}_i^{trans} &= \text{Flip}_x(\mathcal{T}_\Theta(\mathbf{P}_i^{right}; \mathbf{T}_i)), \\
\mathbf{P}_i^{sym} &= \text{Concat}_x(\mathbf{P}_i^{trans}, \mathbf{P}_i^{right}),
\end{aligned}
\tag{2}
$$

in which Θ is the weights of STN; $\mathbf{T}_i$ is the affine transformation matrix; $\mathcal{T}_\Theta$ indicates the affine transformation; Flip_x represents horizontal flipping; and Concat_x stands for concatenation along the x dimension. In Eq. 2, $\mathbf{P}_i^{right}$ can be replaced with $\mathbf{P}_i^{left}$ by swapping the order of the inputs of Concat_x. However, our experiments in §6.3 show that $\mathbf{P}_i^{right}$ performs slightly better than $\mathbf{P}_i^{left}$. For each input $\mathbf{F}_i$ ($i \in \{1, 2, 3, 4\}$), we compute the corresponding $\mathbf{P}_i^{sym}$ using Eq. 2. Using the SPE $\mathbf{P}_i^{sym}$, we recalibrate the input feature map through

$$
\mathbf{F}_i^{recalib} = \mathbf{F}_i + \mathbf{P}_i^{sym}.
\tag{3}
$$

The output $\mathbf{F}_i^{recalib}$ will facilitate the calculation of the subsequent SymAttention.

Micro designs of STN. The spatial transformation in Eq. 2 is conditional on both the input feature and positional encoding. We feed the input feature $\mathbf{F}_i$ into a small network of STN to predict the affine transformation matrix $\mathbf{T}_i$, which is then employed for the affine transformation of the one-side positional encoding $\mathbf{P}_i^{right}$. The small network includes two alternating max-pooling and Conv-ReLU layers. Then, a flattening operation is carried out on the spatial dimension, followed by a multilayer perceptron (MLP) to predict the affine matrix. We initialize the MLP to ensure that the affine transformation with the initial affine matrix is equivalent to an identical mapping.

4.2.2 Symmetric Search Attention

Self-attention has gained popularity in various fields [64]–[67] due to its ability to learn relationships among elements

within a sequence or image [35], [36]. In medical image analysis, self-attention has been applied to identify relevant features in images and enhance disease detection. However, classical self-attention performs global relationship modeling by calculating attention weights for each reference location, which fuses features from all locations. This approach may not be optimal for CTD with CXR images. Specifically, natural images can be captured in various scenarios and contain various objects and elements, so global relationship modeling is beneficial for the understanding of the entire scene. However, CXR images only depict the human chest in a single scenario, and the difference among various CXR images is often limited to the presence of elusive abnormity regions. Therefore, global relationship modeling may be *redundant* for CXR images, limiting the capacity of self-attention to learn relevant relationships for enhancing feature representation. This is because it is challenging for a neural network to automatically identify a few relevant locations out of thousands of redundant locations. For instance, in our experiments, we observe that the DETR detection framework [49] can not converge when used to discriminate indistinguishable TB features in CTD.

To tackle this challenge, we propose SymAttention, which leverages the bilateral symmetry property to aid self-attention in identifying relevant locations in CXR images. As previously mentioned, radiologists can diagnose TB by comparing the bilaterally symmetric locations of the two sides of the lungs. Consequently, the relevant locations for each reference location in CXR images are the bilaterally symmetric locations. Inspired by this, SymAttention searches for features in a symmetrical pattern across the left and right lungs, allowing each reference location only attends to the locations around the bilaterally symmetric location of the reference location. Given the feature map $\mathbf{F}_i^{recalib}$, we first select a small set of key sampling locations, following Deformable DETR [68]. Let K denote the number of selected locations, and M denote the number of heads in the self-attention calculation. The coordinate shifts of the selected locations can be learned by

$$\Delta \mathbf{p}_i^x = \mathbf{W}_x^{pos} \mathbf{F}_i^{recalib}, \qquad \Delta \mathbf{p}_i^y = \mathbf{W}_y^{pos} \mathbf{F}_i^{recalib}, \qquad (4)$$

in which $\mathbf{W}_x^{pos}, \mathbf{W}_y^{pos} \in \mathbb{R}^{(M \times K) \times C}$ are trainable parameter matrices. The attention $\mathbf{A}_i$ and value $\mathbf{F}_i^v$ are simply calculated using

$$\begin{aligned} \mathbf{A}_i &= \text{Softmax}(\text{Reshape}(\mathbf{W}^{att} \mathbf{F}_i^{recalib})), \\ \mathbf{F}_i^v &= \mathbf{W}^{value} \mathbf{F}_i^{recalib} \end{aligned} \qquad (5)$$

where $\mathbf{W}^{att} \in \mathbb{R}^{(M \times K) \times C}$, $\mathbf{W}^{value} \in \mathbb{R}^{C \times C}$ are trainable parameter matrices and the softmax function is performed along the dimension of K. Then, we reshape $\mathbf{F}_i^v$ like

$$\mathbf{F}_i^v \in \mathbb{R}^{C \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}} \rightarrow \mathbf{F}_i^v \in \mathbb{R}^{M \times \frac{C}{M} \times \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}}. \quad (6)$$

Next, SymAttention can be formulated as

$$\mathbf{F}_i^{att} = \text{Concat}_{m=1}^{M} \left(\sum_{k=1}^{K} (\mathbf{A}_i[m,k] \cdot \mathbf{F}_i^v[m,:,\mathbf{p}_i^y + \Delta \mathbf{p}_i^y[m,k], \right.$$
$$\left. \frac{W}{2^{i+1}} - (\mathbf{p}_i^x + \Delta \mathbf{p}_i^x[m,k]) + 1]) \right), \quad (7)$$

in which $\text{Concat}_{m=1}^{M}$ means to concatenate all the results generated by setting m from 1 to M. The term with the wavy underline projects the sampled locations onto the bilaterally symmetric locations by taking the vertical centerline as the line of symmetry, which is the core of the proposed SymAttention. Finally, to ease optimization, a residual connection is connected, followed by an MLP:

$$\hat{\mathbf{F}}_i^{att} = \mathbf{W}^{proj} \mathbf{F}_i^{att} + \mathbf{F}_i, \qquad \hat{\mathbf{F}}_i = \text{MLP}(\hat{\mathbf{F}}_i^{att}) + \hat{\mathbf{F}}_i^{att}, \quad (8)$$

where we have $\mathbf{W}^{proj} \in \mathbb{R}^{C \times C}$ and $\hat{\mathbf{F}}_i$ is the enhanced output as in §4.1.

In Eq. 4 - Eq. 8, each reference location attends to a small set of key sampling locations around the bilaterally symmetric location of the reference location, rather than just the symmetric location. The key sampling locations are automatically set in a learning way. This ensures the receptive field when comparing the appearance of the left and right sides of the lungs. In our experiments, we have observed that the learned coordinate shifts $\Delta \mathbf{p}_i^x$ usually fall within a range of 10% of the width of the corresponding feature map, without any constraints. Consequently, for points situated far from the vertical centerline, they will search points on the symmetric side for feature aggregation. For points in proximity to the vertical centerline, they usually do not correspond to lung regions (see Fig. 6), and this does not impact our assumption of symmetric searching for CTD as we focus on detecting TB in lung regions. In this paper, we empirically set $M = 8$ and $K = 4$. Suppose $N = \frac{H}{2^{i+1}} \times \frac{W}{2^{i+1}}$, and the computational complexity can be expressed as $\mathcal{O}(NC^2)$. Thus, SymAttention is very efficient and flexible for application to the feature pyramid $\mathbf{F}$.

4.3 TB Diagnosis Heads

In §4.1, we mention that there are two TB diagnosis heads: the TB infection area detection head and the CXR image classification head. In this section, we introduce them in detail. The detection head is based on RetinaNet [69], a well-known one-stage object detector, consisting of two branches for bounding box classification and location regression. In contrast to object detection for natural images, where each bounding box covers an object, each bounding box in our system is designed to cover a TB infection area. The detection head learns to detect TB with *two categories*: active TB and latent TB. During clinical TB screening, most CXR cases do not have TB infections, making it easy for the detection head to introduce false positives. To tackle this challenge, we add a CXR image classification head to conduct simultaneous CXR image classification and TB infection area detection. We discard the detected TB areas if a CXR image is classified as non-TB. For simplicity, we stack several convolutions with pooling operations for the classification head. There are five sequential convolution layers, each with 512 output channels and ReLU activation. We then adopt global average pooling to obtain a global feature vector, followed by a fully connected layer with 3 output neurons for classification into *three categories*: healthy, sick but non-TB, and TB.

4.4 Two-stage Training Diagram

Our SymFormer framework consists of two heads designed for CXR image classification and TB infection area detection,

respectively. In clinical settings, the number of non-TB cases significantly outweighs the number of TB cases. Directly training the infection area detection head with non-TB cases would result in an excessive number of pure background supervisions. Therefore, simultaneous training of the classification and detection heads is suboptimal. Additionally, CXR images solely depict structures and organs in the chest, unlike natural images that have complex and diverse backgrounds. If we first train the backbone network and the classification head, the backbone network for feature extraction would become overfitted and would not generalize well to infection area detection. Furthermore, image classification mainly focuses on global features, while infection area detection requires fine-grained features for TB area localization. As a result, training image classification first is also suboptimal. Our proposed approach entails training the backbone network and the infection area detection head initially using only TB CXR images. Then, we employ all CXR images to train the classification head by freezing the backbone network and the detection head. This training strategy benefits from more specific bounding box annotations provided by the detection annotations, which mitigates the risk of overfitting. The fine-grained features learned through the infection area detection can also be easily transferred to CXR image classification.

5 EXPERIMENTAL SETUP

In this section, we first elaborate on the implementation details for the proposed SymFormer in §5.1. Subsequently, we introduce several baseline models for CTD in §5.2 and discuss the evaluation metrics used for CTD in §5.3.

5.1 Implementation Details

Our implementation of SymFormer is done using PyTorch [70] and the open-source `mmdetection` framework [71]. The training of the first stage uses TB CXR images in the TBX11K `trainval` (train + val) set, while the training of the second stage not only uses all TBX11K `trainval` CXR images but also the random half of the MC [25] and Shenzhen [25] datasets as well as the training sets of the DA [6] and DB [6] datasets. The other half of the MC [25] and Shenzhen [25] datasets as well as the test sets of the TBX11K, DA [6] and DB [6] datasets are used to evaluate the performance of CXR image classification. We set the number of FPN feature channels, denoted as C, to 256, consistent with RetinaNet [69]. Other settings also follow those in RetinaNet. For the training of the first stage, we use a batch size of 8 and train for 50 epochs for Deformable DETR-based models [68], while 24 epochs for other models. For the training of the second stage, we utilize a batch size of 8 and train for 12 epochs across all models. We adopt the AdamW optimizer for Deformable DETR-based models and the SGD optimizer for other models. To augment the data, we use random flipping. We resize both the CXR images used for training and testing to 512×512. All experiments are conducted using 2 TITAN XP GPUs. Please refer to our code for more details.

5.2 Baseline Models

As discussed in §4.1, incorporating an image classification head can significantly reduce the false positives of detection in clinical TB screening. However, existing object detectors do not consider background images and often disregard images without bounding-box objects [69], [72]–[75]. Using these detectors directly for CTD leads to numerous false positives due to the large number of non-TB CXR images in clinical practice. To address this issue, we introduce a classification head to enable simultaneous CXR image classification and TB infection area detection, where the CXR image classification results are used to filter out the false positives of detection.

To achieve this, we reformulate several well-known object detectors, including SSD [72], RetinaNet [69], Faster R-CNN [74], FCOS [73], and Deformable DETR [68] for simultaneous CXR image classification and TB infection area detection. Specifically, we add the same image classification head as used in our SymFormer to these detectors, after the final layer of their backbone networks, *e.g.*, *conv5_3* for VGGNet-16 [76] and *res5c* for ResNet-50 [62]. The image classification head learns to classify CXR images into three categories: healthy, sick but non-TB, and TB, while the TB detection head learns to detect TB with two categories: active TB and latent TB. The training of existing detectors follows the two-stage training diagram described in §4.4.

5.3 Evaluation Metrics

CXR image classification. We continue by introducing the evaluation metrics for the CTD task. In CXR image classification, the goal is to classify each CXR image into one of three categories: healthy, sick but non-TB, and TB. To assess the classification results, we utilize the following evaluation metrics:

- Accuracy: This metric measures the percentage of CXR images that are correctly classified across all three categories.
- Area Under Curve (AUC): AUC computes the area under the Receiver Operating Characteristic (ROC) curve. The ROC curve plots the true positive rate against the false positive rate for the TB class.
- Sensitivity: Sensitivity quantifies the percentage of TB cases that are accurately identified as TB. It represents the recall for the TB class.
- Specificity: Specificity determines the percentage of non-TB cases that are correctly identified as non-TB, encompassing both the healthy and sick but non-TB classes. It represents the recall for the non-TB class.
- Average Precision (AP): AP calculates the precision for each class and takes the average across all classes. It provides an overall measure of precision.
- Average Recall (AR): AR computes the recall for each class and averages the values across all classes. It provides an overall measure of recall.
- Confusion matrix: The confusion matrix reports the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). For a better view, we report the ratios of TP, TN, FP, and FN in relation to the total number of test CXR images.

TABLE 3
CXR image classification results (%) in terms of accuracy, AUC, sensitivity, specificity, AP, and AR on the TBX11K test data. The "Backbone" column indicates the specific backbone network used.

Methods	Backbones	Accuracy	AUC (TB)	Sensitivity	Specificity	Ave. Prec. (AP)	Ave. Rec. (AR)
SSD [72]	VGGNet-16	84.7	93.0	78.1	89.4	82.1	83.8
RetinaNet [69]	ResNet-50 w/ FPN	87.4	91.8	81.6	89.8	84.8	86.8
Faster R-CNN [74]	ResNet-50 w/ FPN	89.7	93.6	91.2	89.9	87.7	90.5
FCOS [73]	ResNet-50 w/ FPN	88.9	92.4	87.3	89.9	86.6	89.2
Deformable DETR [68]	ResNet-50 w/ FPN	91.3	97.6	89.2	95.3	89.8	91.0
SymFormer w/ Deformable DETR	ResNet-50 w/ FPN	94.3	98.5	87.3	**97.3**	93.2	93.2
SymFormer w/ RetinaNet	ResNet-50 w/ FPN	94.5	98.9	91.0	96.8	93.3	94.0
SymFormer w/ RetinaNet	P2T-Small w/ FPN	**94.6**	**99.1**	**92.1**	96.7	**93.4**	**94.2**

TABLE 4
CXR image classification results (%) in terms of the F_1 score and confusion matrix on the TBX11K test data, as well as the number of FLOPs, the number of parameters, and FPS of each model. "#Total" denotes the total number of test CXR images. We test FPS on a single TITAN XP GPU. For the ground truths, the ratio of positives (TP + FN) is 19.6%, and the ratio of negatives (TN + FP) is 80.4%.

Methods	Backbones	#FLOPs	#Params	FPS	F_1 score ↑	TP/#Total ↑	TN/#Total ↑	FP/#Total ↓	FN/#Total ↓
SSD [72]	VGGNet-16	90.58	38.69	32.9	70.5	15.3	71.9	8.5	4.3
RetinaNet [69]	ResNet-50 w/ FPN	55.41	48.97	35.3	73.1	16.0	72.2	8.2	3.6
Faster R-CNN [74]	ResNet-50 w/ FPN	66.27	53.98	30.3	78.5	17.9	72.3	8.1	1.7
FCOS [73]	ResNet-50 w/ FPN	53.33	44.69	39.9	76.3	17.1	72.3	8.1	2.5
Deformable DETR [68]	ResNet-50 w/ FPN	54.07	52.67	23.0	85.6	17.5	76.6	3.8	2.1
SymFormer w/ Deformable DETR	ResNet-50 w/ FPN	54.08	52.69	22.5	87.9	17.1	**78.2**	**2.2**	2.5
SymFormer w/ RetinaNet	ResNet-50 w/ FPN	59.14	50.03	24.3	89.0	17.8	77.8	2.6	1.8
SymFormer w/ RetinaNet	P2T-Small w/ FPN	55.46	45.10	17.9	**89.6**	**18.1**	77.7	2.7	**1.5**

- F_1 score: This metric is the harmonic mean of precision and recall, thus symmetrically representing both in a single value. It can be calculated by 2TP/(2TP + FP + FN).

These metrics enable the evaluation of the CXR image classification quality from various perspectives.

TB infection area detection. For the evaluation of TB detection, we utilize the average precision of the bounding box metric (AP^{bb}) proposed by the COCO benchmark [77]. AP^{bb} is widely used as the primary detection metric in the vision community [38], [69], [73], [78]. The default AP^{bb} is computed by averaging over IoU (intersection-over-union) thresholds ranging from 0.5 to 0.95 with a step size of 0.05. Additionally, we report AP_{50}^{bb}, which represents AP^{bb} at an IoU threshold of 0.5. To provide insights into the detection performance for different types of TB, we present evaluation results separately for active TB and latent TB, excluding uncertain TB CXR images. We also report category-agnostic TB detection results, where the TB categories are disregarded, to describe the detection of all TB areas. In this case, uncertain TB CXR images are included. Furthermore, we introduce two evaluation modes: i) utilizing all CXR images in the TBX11K test set, and ii) considering only TB CXR images in the TBX11K test set. By employing these metrics, we can comprehensively analyze the performance of CTD systems from various useful perspectives.

6 EXPERIMENTAL RESULTS

In this section, we present the results for CXR image classification in §6.1, followed by the results for TB infection area detection in §6.2. Subsequently, we visualize detection results and the learned deep features in §6.4. Lastly, we conduct ablation studies in §6.3 to gain a better understanding of the proposed SymFormer.

6.1 CXR Image Classification

We summarize the evaluation results for CXR image classification on the TBX11K test set in Table 3 and Table 4. All methods adopt pretraining models from ImageNet [32] for initialization. We report the results of the proposed SymFormer integrated with RetinaNet [69] and Deformable DETR [68] as the base methods. As can be observed from both Table 3 and Table 4, incorporating SymFormer into RetinaNet [69] and Deformable DETR [68] leads to significant performance improvements for RetinaNet and Deformable DETR, respectively. SymFormer with Deformable DETR achieves a specificity of 97.3%, indicating that 2.7 out of 100 non-TB CXR images will be misclassified as TB. The *default model* we employ is SymFormer with RetinaNet, which exhibits slightly lower performance than SymFormer with Deformable DETR but outperforms the latter by a significant margin in object detection, as demonstrated in §6.2. Furthermore, in terms of accuracy, all methods greatly outperform radiologists who achieve an accuracy of 84.8% as in §3.2. This emphasizes the promising potential of deep-learning-based CTD as a research field.

In Table 3, we observe that the difference between SymFormer and the baseline models in terms of sensitivity is much smaller than the difference in terms of specificity. Specially, Faster R-CNN [74] achieves an impressively high sensitivity rate of 91.2%, but it lags significantly behind SymFormer in other performance metrics. To explain this phenomenon, we refer to Table 4 and discover that the baseline models tend to make more positive predictions (TP

TABLE 5
CXR image classification results (%) in terms of accuracy, AUC, sensitivity, specificity, AP, and AR on the DA+DB test data [6].

Methods	Backbones	Accuracy	AUC (TB)	Sensitivity	Specificity	Ave. Prec. (AP)	Ave. Rec. (AR)
SSD [72]	VGGNet-16	51.0	53.8	**100.0**	1.9	75.3	51.0
RetinaNet [69]	ResNet-50 w/ FPN	50.0	50.0	**100.0**	0.0	25.0	50.0
Faster R-CNN [74]	ResNet-50 w/ FPN	50.0	51.9	**100.0**	0.0	25.0	50.0
FCOS [73]	ResNet-50 w/ FPN	50.0	52.1	**100.0**	0.0	25.0	50.0
Deformable DETR [68]	ResNet-50 w/ FPN	68.6	69.7	84.3	52.9	70.7	68.6
SymFormer w/ Deformable DETR	ResNet-50 w/ FPN	82.4	78.4	86.3	78.4	82.6	82.4
SymFormer w/ RetinaNet	ResNet-50 w/ FPN	78.4	74.7	90.2	66.7	80.1	78.4
SymFormer w/ RetinaNet	P2T-Small w/ FPN	**84.3**	**89.4**	84.3	**84.3**	**84.3**	**84.3**

TABLE 6
CXR image classification results (%) in terms of accuracy, AUC, sensitivity, specificity, AP, and AR on the MC+Shenzhen test data [25].

Methods	Backbones	Accuracy	AUC (TB)	Sensitivity	Specificity	Ave. Prec. (AP)	Ave. Rec. (AR)
SSD [72]	VGGNet-16	50.8	50.4	**100.0**	3.4	74.9	51.7
RetinaNet [69]	ResNet-50 w/ FPN	49.3	49.7	**100.0**	0.5	74.6	50.3
Faster R-CNN [74]	ResNet-50 w/ FPN	49.0	49.5	**100.0**	0.0	24.5	50.0
FCOS [73]	ResNet-50 w/ FPN	48.8	49.0	99.5	0.0	24.4	49.7
Deformable DETR [68]	ResNet-50 w/ FPN	81.3	83.5	92.9	70.1	83.0	81.5
SymFormer w/ Deformable DETR	ResNet-50 w/ FPN	82.0	84.7	89.3	75.0	82.7	82.1
SymFormer w/ RetinaNet	ResNet-50 w/ FPN	82.8	86.3	91.8	74.0	83.8	82.9
SymFormer w/ RetinaNet	P2T-Small w/ FPN	**85.8**	**87.4**	93.4	**78.4**	**86.6**	**85.9**

TABLE 7
TB infection area detection results (%) on our TBX11K test set. The "Test Data" column specifies whether the evaluation was performed using all CXR images in the test set or only TB CXR images in the test set. The "Backbone" column indicates the specific backbone network used.

Methods	Test Data	Backbones	Category-agnostic TB		Active TB		Latent TB	
			AP^{bb}_{50}	AP^{bb}	AP^{bb}_{50}	AP^{bb}	AP^{bb}_{50}	AP^{bb}
SSD [72]	ALL	VGGNet-16	52.3	22.6	50.5	22.8	8.1	3.2
RetinaNet [69]		ResNet-50 w/ FPN	52.1	22.2	45.4	19.6	6.2	2.4
Faster R-CNN [74]		ResNet-50 w/ FPN	57.3	22.7	53.3	21.9	9.6	2.9
FCOS [73]		ResNet-50 w/ FPN	46.6	18.9	40.3	16.8	6.2	2.1
Deformable DETR [68]		ResNet-50 w/ FPN	51.7	22.0	48.9	21.2	7.1	1.9
SymFormer w/ Deformable DETR		ResNet-50 w/ FPN	57.0	23.3	52.1	22.7	7.1	2.0
SymFormer w/ RetinaNet		ResNet-50 w/ FPN	68.0	29.5	62.0	**27.3**	**13.3**	**4.4**
SymFormer w/ RetinaNet		P2T-Small w/ FPN	**70.4**	**30.0**	**63.6**	26.9	11.4	4.3
SSD [72]	Only TB	VGGNet-16	68.3	28.7	63.7	28.0	10.7	4.0
RetinaNet [69]		ResNet-50 w/ FPN	69.4	28.3	61.5	25.3	10.2	4.1
Faster R-CNN [74]		ResNet-50 w/ FPN	63.4	24.6	58.7	23.7	9.6	2.8
FCOS [73]		ResNet-50 w/ FPN	56.3	22.5	47.9	19.8	7.4	2.4
Deformable DETR [68]		ResNet-50 w/ FPN	57.4	24.2	54.5	23.5	7.6	2.3
SymFormer w/ Deformable DETR		ResNet-50 w/ FPN	60.8	24.5	55.2	23.8	9.2	2.6
SymFormer w/ RetinaNet		ResNet-50 w/ FPN	73.4	31.5	67.1	**29.2**	**14.7**	**4.8**
SymFormer w/ RetinaNet		P2T-Small w/ FPN	**75.7**	**32.1**	**68.9**	28.9	13.0	4.7

+ FP) and fewer negative predictions (TN + FN). In simpler terms, the baseline models are inclined to classify a test CXR image as positive, potentially due to their limited ability to learn high-quality TB-related features. When we evaluate all models on other public datasets without retraining, as shown in Table 5 and Table 6, we can see that baseline models even achieve a sensitivity rate of 100.0% and a specificity rate of 0. This further confirms our hypothesis that baseline models tend to classify CXR images as positive. Considering this perspective, the F_1 score in Table 4 provides a better representation of a model's overall performance, as it symmetrically combines precision and recall.

In Table 4, we also report the number of Floating-Point Operations (FLOPs), the number of parameters, and Frames Per Second (FPS) for each model. From the comparison between the original and SymFormer-enhanced Deformable DETR [68], we can see that SymFormer exhibits similar FLOPs, parameters, and running speed as Deformable DETR. This is straightforward as SymFormer only introduces negligible computations to Deformable DETR. When integrated with RetinaNet [69] using the ResNet-50 [62] backbone, SymFormer achieves 24.3 fps, making it a practical choice for deployment in real-world scenarios. If additional computational resources are available, SymFormer can also utilize P2T-Small [38] as the backbone, offering enhanced diagnostic performance and a speed of 17.9 fps.

In addition, we assess CXR image classification on public datasets using the above-trained models without fine-tuning. The results are presented in Table 5 for DA+DB test data [6] and Table 6 for MC+Shenzhen test data [25].
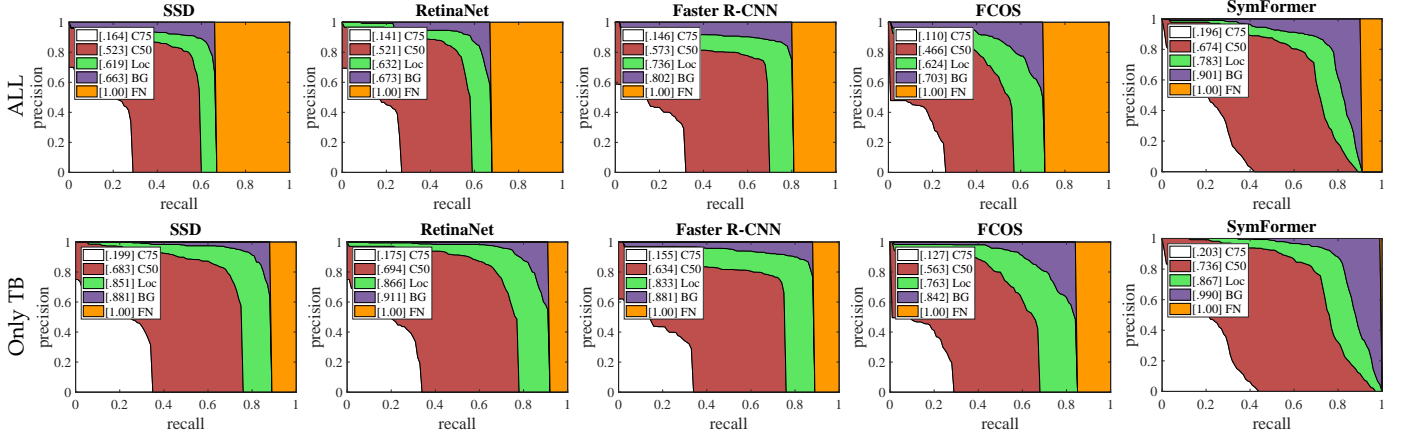
Fig. 5. **Error analyses of category-agnostic TB area detection using baseline models and SymFormer w/ RetinaNet.** The first row is evaluated using all CXR images, while the second row only uses TB CXR images. C50/C75: PR curves under IoU thresholds of 0.5/0.75. Loc: the PR curve under the IoU threshold of 0.1. BG: removing background false positives. FN: removing other errors caused by undetected targets (false negatives). SymFormer largely outperforms other methods in all metrics, *e.g.*, obtaining a remarkable 99% BG score when only using TB CXR images.

Notably, SSD [72], RetinaNet [69], Faster R-CNN [74], and FCOS [73] achieve a sensitivity rate of approximately 100.0% and a specificity rate of around 0. As discussed earlier, this suggests that baseline models struggle to learn robust feature representations for CTD, often misclassifying test CXR images as TB cases. Deformable DETR [68] demonstrates a degree of generalization to public datasets but falls short in comparison to the proposed SymFormer. The strong performance of SymFormer underscores its exceptional generalization capability.

6.2 TB Infection Area Detection

We proceed by presenting the results for TB infection area detection. As discussed in §5.3, we report the performance for both the entire TBX11K test set and a subset consisting only of TB CXR images. Evaluating the performance using only TB CXR images allows for precise detection analysis since non-TB CXR images do not contain target TB infection areas. Conversely, evaluating using all CXR images incorporates the influence of false positives in non-TB CXR images. To ensure accurate evaluation using all CXR images, we *discard* all predicted boxes in CXR images that are classified as non-TB by the CXR image classification head. However, it is important to note that this filtering process is not applicable when evaluating using only TB CXR images.

The results for TB infection area detection are presented in Table 7. It is evident that both SymFormer with Deformable DETR and SymFormer with RetinaNet demonstrate significant improvements over their respective base methods, Deformable DETR [68] and RetinaNet [69]. Interestingly, SymFormer with RetinaNet outperforms SymFormer with Deformable DETR by a considerable margin, indicating that SymFormer is better suited for integration with the RetinaNet framework. As a result, we select SymFormer with RetinaNet as our default model for CTD. It is worth noting that all methods struggle with accurately detecting latent TB areas. However, the evaluation results for category-agnostic TB are better than those for active TB, indicating that many latent TB targets are correctly located but mistakenly classified as active TB. We attribute this to

TABLE 8
Ablation study for TB infection area detection on our TBX11K validation set. We only use TB CXR images to evaluate category-agnostic TB area detection. The "Symmetry of SPE" column indicates whether SPE transfers the right side of positional encoding to the left side, or vice versa. APE: absolute positional encoding; RPE: relative positional encoding.

Attention	Positional Encoding	Symmetry of SPE	AP_{50}^{bb}	AP^{bb}
No	No	-	72.7	31.0
Vanilla	APE	-	73.4	30.6
Vanilla	RPE	-	72.7	29.7
Vanilla	SPE w/o STN	left → right	74.0	30.5
Vanilla	SPE w/o STN	right → left	74.3	30.8
Vanilla	SPE	left → right	75.1	30.4
Vanilla	SPE	right → left	75.7	29.6
SymAttention	APE	-	74.9	30.0
SymAttention	RPE	-	73.6	29.1
SymAttention	SPE w/o STN	left → right	75.3	31.4
SymAttention	SPE w/o STN	right → left	75.5	30.7
SymAttention	SPE	left → right	76.3	30.9
SymAttention	SPE	right → left	**76.6**	**31.7**

the limited number of latent TB CXR images in the TBX11K dataset, where only 212 CXR images depict latent TB compared to 924 CXR images depicting active TB. Therefore, future research should address this data imbalance issue and focus on improving the detection of latent TB areas. Furthermore, we observe that the performance in terms of AP_{50}^{bb} is generally superior to that of AP^{bb}. This suggests that while detection models are capable of identifying the target regions, their localization accuracy is often not very precise. We argue that locating TB bounding box regions differs significantly from locating regions of natural objects. Even experienced radiologists find it challenging to precisely pinpoint TB regions. Consequently, AP_{50}^{bb} is more crucial than AP^{bb} since predicted boxes with an IoU of 0.5 with target TB areas are sufficient to assist radiologists in identifying TB infection areas.

In Fig. 5, we present the precision-recall (PR) curves for the detection error analyses, focusing on category-agnostic TB detection. It is evident that all methods exhibit substantial improvements when transitioning from an IoU thresh-

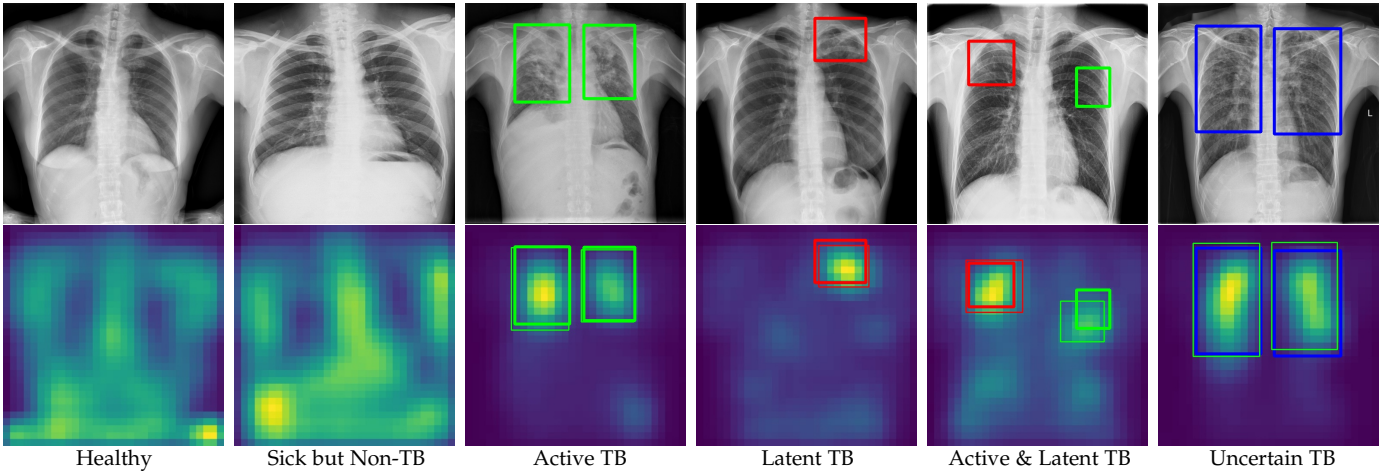| Healthy | Sick but Non-TB | Active TB | Latent TB | Active & Latent TB | Uncertain TB |

Fig. 6. **Visualization of the learned deep features from CXR images using SymFormer w/ RetinaNet.** We randomly select CXR images from the TBX11K test set, and for each class mentioned in Table 2, we provide one example. In each example, the infection areas of active TB, latent TB, and uncertain TB are indicated by boxes colored in green, red, and blue, respectively. The ground-truth boxes are displayed with thick lines, while the detected boxes are shown with thin lines.

TABLE 9
Evaluation results (%) of the 4-fold cross-validation. The used model is our SymFormer w/ RetinaNet using ResNet-50 w/ FPN as the backbone. We split the TBX11K `trainval` set into 4 folds, each of which has a similar class distribution. "#Total" denotes the total number of test CXR images in each fold. We only use TB CXR images to evaluate category-agnostic TB area detection.

Fold	Accuracy	AUC	Sensitivity	Specificity	AP	AR	F_1 score	TP/#Total	TN/#Total	FP/#Total	FN/#Total	AP_{50}^{bb}	AP^{bb}
1	94.7	98.9	92.2	97.5	91.1	94.6	87.6	11.0	85.8	2.2	0.9	74.7	31.7
2	95.2	98.9	92.6	97.6	91.7	95.0	88.1	11.1	86.0	2.1	0.9	75.2	30.4
3	94.6	99.1	92.9	96.9	90.5	94.6	86.4	11.1	85.3	2.7	0.8	74.1	29.5
4	95.1	99.3	92.9	97.3	91.2	94.7	87.4	11.1	85.7	2.4	0.8	75.4	33.3

old of 0.75 to 0.5. This indicates that the performance of all methods is particularly challenged at higher IoU thresholds due to their limited object localization capabilities. Comparing the results obtained using all CXR images with those using only TB CXR images, we observe that the region labeled as "FN" (false negatives) is larger when evaluating using all CXR images. This suggests that the filtering process based on image classification disregards many correctly detected TB areas, despite its effectiveness in improving overall detection performance. Importantly, the "FN" region for SymFormer is significantly smaller than that of other methods, highlighting its superior ability to detect fewer false negatives. Regardless of whether all CXR images or only TB CXR images are utilized, SymFormer consistently exhibits higher PR curves for IoU thresholds of 0.75, 0.5, and 0.1. By considering the results of both image classification and TB infection area detection, we can confidently conclude that the proposed SymFormer achieves state-of-the-art performance and serves as a strong baseline for future research in the field of CTD.

6.3 Ablation Study

In this part, we first carry out ablation studies to investigate the effectiveness of the proposed modules. Specifically, we train the models using the training set of our TBX11K dataset and evaluate them on the validation set. The results are presented in Table 8. The baseline model is RetinaNet [69], which corresponds to the first model in Table 8 and does not incorporate any attention or positional encoding.

The term "vanilla attention" refers to the deformable attention employed in Deformable DETR [68]. We utilize well-established implementations for both absolute positional encoding [35], [36] (as described in Eq. 1) and relative positional encoding [79]. As specified in Eq. 2, the default version of SPE transfers the right side of the positional encoding to the left side. Here, we also evaluate the performance when transferring the left side to the right side.

Based on the discussions in §6.2, the AP_{50}^{bb} metric is deemed sufficient for measuring the effectiveness of a model in assisting radiologists with identifying TB infection areas. As evident from Table 8, relative positional encoding achieves inferior performance compared to absolute positional encoding, leading us to construct our SPE using absolute positional encoding. Besides, the addition of absolute positional encoding and any form of attention to RetinaNet [69] yields significant improvements in detection performance. Furthermore, across all types of positional encoding, our proposed SymAttention consistently outperforms deformable attention, showcasing its superiority in learning distinctive representations for CTD. Notably, even without STN, the proposed SPE consistently achieves superior performance compared to both absolute positional encoding and relative positional encoding. The inclusion of STN further enhances the performance of SPE, confirming its effectiveness. Therefore, our investigation into symmetric abnormality search in CTD has yielded successful results. In addition, we can observe that, for the symmetry of SPE, the transfer of positional encoding from right to left, as opposed

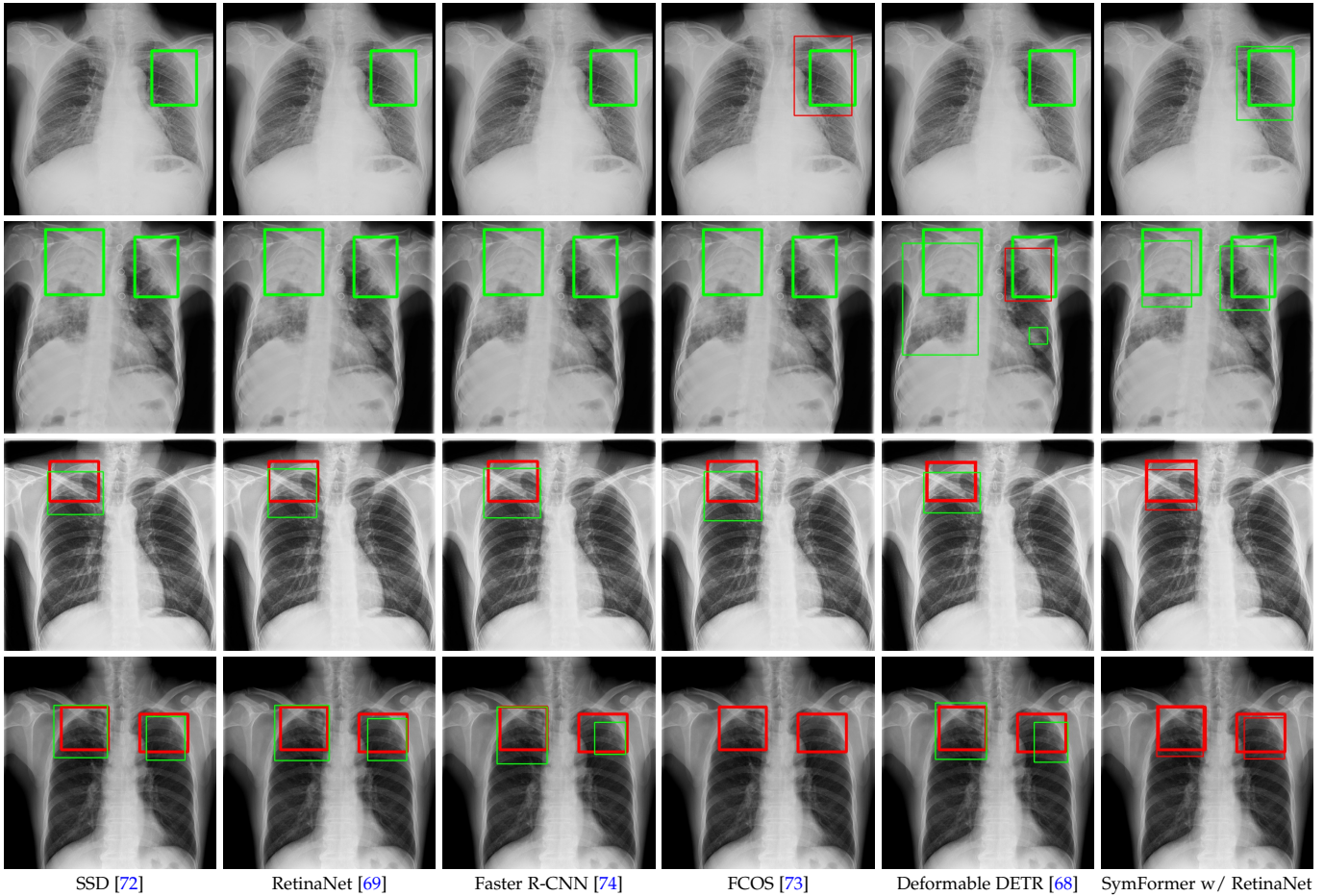| SSD [72] | RetinaNet [69] | Faster R-CNN [74] | FCOS [73] | Deformable DETR [68] | SymFormer w/ RetinaNet |

Fig. 7. **Qualitative comparison between the proposed SymFormer and baseline methods.** In each example, the infection areas of active TB and latent TB are indicated by green boxes and red boxes, respectively. The ground-truth boxes are displayed with thick lines, while the detected boxes are shown with thin lines. For all examples, SymFormer can detect all TB infection areas with true categories.

to left to right, slightly outperforms. Thus, we transfer the positional encoding from right to left by default.

To demonstrate the robustness of the proposed Sym-Former, we perform a 4-fold cross-validation. We partition the TBX11K `trainval` data into four folds, ensuring that each fold maintains a similar class distribution. During each trial, we train SymFormer on three of these folds and evaluate its performance on the remaining fold. The evaluation results of the four trials are displayed in Table 9. As can be observed, the results across these different trials are remarkably consistent, affirming the robustness of SymFormer.

6.4 Visualization

To gain insights into the learning process of deep neural networks on CXR images, we visualize the feature map of SymFormer w/ RetinaNet at a scale of 1/32. To achieve this, we employ principal component analysis (PCA) to reduce the channels of the feature map to a single channel. The resulting single-channel map is then converted into a heat map for visualization purposes. The visualization of the learned features, along with the corresponding detection results, are presented in Fig. 6. Upon analysis, we observe that the visualization of healthy cases exhibits irregular feature patterns, indicating the absence of significant abnormalities. In contrast, the visualization of sick but non-TB cases displayed some discernible highlights, potentially representing the presence of lesions. For TB cases, the highlights in the visualization map align well with the annotated TB infection areas, thereby indicating the effectiveness of the proposed SymFormer in learning deep features for TB area detection. Furthermore, in Fig. 7, we offer qualitative comparisons between the proposed SymFormer and the baseline models for TB infection area detection. As evident, SymFormer consistently delivers superior qualitative detection results.

7 CONCLUSION

Early diagnosis plays a crucial role in effectively treating and preventing tuberculosis (TB), a prevalent infectious disease worldwide. However, TB diagnosis remains a significant challenge, particularly in resource-constrained communities and developing countries. The conventional gold standard test for TB necessitates a BSL-3 laboratory and is a time-consuming process, taking several months to provide definitive results, making it impractical in many settings. Deep learning has shown promising advancements in various domains, prompting researchers to explore its potential in computer-aided TB diagnosis (CTD). Nonetheless, the lack of annotated data has hindered the progress of deep learning in this field. To address this limitation, we introduce TBX11K, a large-scale TB dataset with bounding box annotations. This dataset not only facilitates the training

of deep neural networks for CTD but also serves as the first dataset specifically designed for TB detection.

In addition to the dataset, we propose a simple yet effective framework called SymFormer for simultaneous CXR image classification and TB infection area detection. Leveraging the *bilateral symmetry property* inherent in CXR images, SymFormer incorporates Symmetric Search Attention (SymAttention) to extract distinctive feature representations. Recognizing that CXR images may not exhibit strict symmetry, we introduce Symmetric Positional Encoding (SPE) to enhance the performance of SymAttention through feature recalibration. Furthermore, to provide a benchmark for CTD research, we introduce evaluation metrics, assess baseline models adapted from existing object detectors, and launch an online challenge. Our experiments demonstrate that SymFormer achieves state-of-the-art performance on the TBX11K dataset, positioning it as a strong baseline for future research endeavors. The introduction of the TBX11K dataset, the SymFormer method, and the CTD benchmark in this study are expected to significantly advance research in the field of CTD, ultimately contributing to improved detection and management of TB worldwide.

ACKNOWLEDGMENTS

REFERENCES

[1] Y. Liu, Y.-H. Wu, Y. Ban, H. Wang, and M.-M. Cheng, "Rethinking computer-aided tuberculosis diagnosis," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 2646–2655.

[2] W. H. Organization, "Global tuberculosis report 2015," https://apps.who.int/iris/handle/10665/191102, 2015.

[3] ——, "Global tuberculosis report 2017," https://apps.who.int/iris/handle/10665/259366, 2017.

[4] ——, "Global tuberculosis report 2020," https://www.who.int/publications/i/item/9789240013131, 2020.

[5] S. Jaeger, A. Karargyris, S. Candemir, L. Folio, J. Siegelman, F. Callaghan, Z. Xue, K. Palaniappan, R. K. Singh, S. Antani *et al.*, "Automatic tuberculosis screening using chest radiographs," *IEEE Trans. Med. Imaging*, vol. 33, no. 2, pp. 233–245, 2013.

[6] A. Chauhan, D. Chauhan, and C. Rout, "Role of gist and phog features in computer-aided diagnosis of tuberculosis without segmentation," *PloS One*, vol. 9, no. 11, p. e112980, 2014.

[7] S. Hwang, H.-E. Kim, J. Jeong, and H.-J. Kim, "A novel approach for tuberculosis screening based on deep convolutional neural networks," in *Medical Imaging: Computer-Aided Diagnosis*, vol. 9785. International Society for Optics and Photonics, 2016, p. 97852W.

[8] N. R. Gandhi, P. Nunn, K. Dheda, H. S. Schaaf, M. Zignol, D. Van Soolingen, P. Jensen, and J. Bayona, "Multidrug-resistant and extensively drug-resistant tuberculosis: a threat to global control of tuberculosis," *The Lancet*, vol. 375, no. 9728, pp. 1830–1843, 2010.

[9] U. Lopes and J. F. Valiati, "Pre-trained convolutional neural networks as feature extractors for tuberculosis detection," *Computers in Biology and Medicine*, vol. 89, pp. 135–143, 2017.

[10] S. Candemir, S. Jaeger, K. Palaniappan, J. P. Musco, R. K. Singh, Z. Xue, A. Karargyris, S. Antani, G. Thoma, and C. J. McDonald, "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," *IEEE Trans. Med. Imaging*, vol. 33, no. 2, pp. 577–590, 2013.

[11] P. Andersen, M. Munk, J. Pollock, and T. Doherty, "Specific immune-based diagnosis of tuberculosis," *The Lancet*, vol. 356, no. 9235, pp. 1099–1104, 2000.

[12] A. Bekmurzayeva, M. Sypabekova, and D. Kanayeva, "Tuberculosis diagnosis using immunodominant, secreted antigens of mycobacterium tuberculosis," *Tuberculosis*, vol. 93, no. 4, pp. 381–388, 2013.

[13] W. H. Organization *et al.*, "Chest radiography in tuberculosis detection: Summary of current WHO recommendations and guidance on programmatic approaches," World Health Organization, Tech. Rep., 2016.

[14] M. Van Cleeff, L. Kivihya-Ndugga, H. Meme, J. Odhiambo, and P. Klatser, "The role and performance of chest X-ray for the diagnosis of tuberculosis: A cost-effectiveness analysis in nairobi, kenya," *BMC Infectious Diseases*, vol. 5, no. 1, p. 111, 2005.

[15] A. Konstantinos, "Testing for tuberculosis," *Australian Prescriber*, vol. 33, no. 1, pp. 12–18, 2010.

[16] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Adv. Neural Inform. Process. Syst.*, 2014, pp. 1988–1996.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.

[19] T. Xu, X.-F. Zhu, and X.-J. Wu, "Learning spatio-temporal discriminative model for affine subspace based visual object tracking," *Visual Intelligence*, vol. 1, no. 1, p. 4, 2023.

[20] Y. Liu, M.-M. Cheng, X. Hu, J.-W. Bian, L. Zhang, X. Bai, and J. Tang, "Richer convolutional features for edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1939–1946, 2019.

[21] Y. Liu, M.-M. Cheng, D.-P. Fan, L. Zhang, J.-W. Bian, and D. Tao, "Semantic edge detection with diverse deep supervision," *Int. J. Comput. Vis.*, vol. 130, no. 1, pp. 179–198, 2022.

[22] J. Li, Z. Li, L. Wei, and X. Zhang, "Machine learning in lung cancer radiomics," *Machine Intelligence Research*, 2023.

[23] Y. Qiu, F. Lin, W. Chen, and M. Xu, "Pre-training in medical data: A survey," *Machine Intelligence Research*, vol. 20, no. 2, pp. 147–179, 2023.

[24] W.-C. Wang, E. Ahn, D. Feng, and J. Kim, "A review of predictive and contrastive self-supervised learning for medical images," *Machine Intelligence Research*, vol. 20, no. 4, pp. 483–513, 2023.

[25] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative Imaging in Medicine and Surgery*, vol. 4, no. 6, pp. 475–477, 2014.

[26] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient N-D image segmentation," *Int. J. Comput. Vis.*, vol. 70, no. 2, pp. 109–131, 2006.

[27] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Progress in Brain Research*, vol. 155, pp. 23–36, 2006.

[28] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *ACM International Conference on Image and Video Retrieval*. ACM, 2007, pp. 401–408.

[29] M.-M. Cheng, Y. Liu, Q. Hou, J. Bian, P. Torr, S.-M. Hu, and Z. Tu, "HFS: Hierarchical feature selection for efficient image segmentation," in *Eur. Conf. Comput. Vis.*, 2016, pp. 867–882.

[30] Y. Liu, P.-T. Jiang, V. Petrosyan, S.-J. Li, J. Bian, L. Zhang, and M.-M. Cheng, "DEL: Deep embedding learning for efficient image segmentation," in *Int. Joint Conf. Artif. Intell.*, 2018, pp. 864–870.

[31] A. Karargyris, J. Siegelman, D. Tzortzis, S. Jaeger, S. Candemir, Z. Xue, K. Santosh, S. Vajda, S. Antani, L. Folio *et al.*, "Combination of texture and shape features to detect pulmonary abnormalities in digital chest X-rays," *International Journal of Computer Assisted Radiology and Surgery*, vol. 11, no. 1, pp. 99–106, 2016.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 248–255.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Adv. Neural Inform. Process. Syst.*, 2012, pp. 1097–1105.

[34] P. Lakhani and B. Sundaram, "Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574–582, 2017.

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Adv. Neural Inform. Process. Syst.*, pp. 5998–6008, 2017.

[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly

et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Int. Conf. Learn. Represent.*, 2021.

[37] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Int. Conf. Mach. Learn.*, 2021, pp. 10 347–10 357.

[38] Y.-H. Wu, Y. Liu, X. Zhan, and M.-M. Cheng, "P2T: Pyramid pooling transformer for scene understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.

[39] Y. Zhang, H. Liu, and Q. Hu, "TransFuse: Fusing transformers and CNNs for medical image segmentation," in *Med. Image. Comput. Comput. Assist. Interv.*, 2021, pp. 14–24.

[40] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "CoTr: Efficiently bridging CNN and transformer for 3D medical image segmentation," in *Med. Image. Comput. Comput. Assist. Interv.*, 2021, pp. 171–180.

[41] Y. Ji, R. Zhang, H. Wang, Z. Li, L. Wu, S. Zhang, and P. Luo, "Multi-compound transformer for accurate biomedical image segmentation," in *Med. Image. Comput. Comput. Assist. Interv.*, 2021, pp. 326–336.

[42] Y. Gao, M. Zhou, and D. N. Metaxas, "UTNet: A hybrid transformer architecture for medical image segmentation," in *Med. Image. Comput. Comput. Assist. Interv.*, 2021, pp. 61–71.

[43] R. Tao, W. Liu, and G. Zheng, "Spine-transformers: Vertebra labeling and segmentation in arbitrary field-of-view spine CTs via 3D transformers," *Med. Image Anal.*, vol. 75, p. 102258, 2022.

[44] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji *et al.*, "TransMIL: Transformer based correlated multiple instance learning for whole slide image classification," *Adv. Neural Inform. Process. Syst.*, pp. 2136–2147, 2021.

[45] S. Park, G. Kim, J. Kim, B. Kim, and J. C. Ye, "Federated split task-agnostic vision transformer for COVID-19 CXR diagnosis," *Adv. Neural Inform. Process. Syst.*, pp. 24 617–24 630, 2021.

[46] Y. Mo, C. Han, Y. Liu, M. Liu, Z. Shi, J. Lin, B. Zhao, C. Huang, B. Qiu, Y. Cui *et al.*, "HoVer-Trans: Anatomy-aware HoVer-Transformer for ROI-free breast cancer diagnosis in ultrasound images," *IEEE Trans. Med. Imaging*, vol. 42, no. 6, pp. 1696–1706, 2023.

[47] M. Bhattacharya, S. Jain, and P. Prasanna, "RadioTransformer: A cascaded global-focal transformer for visual attention–guided disease classification," in *Eur. Conf. Comput. Vis.*, 2022, pp. 679–698.

[48] S. Park, G. Kim, Y. Oh, J. B. Seo, S. M. Lee, J. H. Kim, S. Moon, J.-K. Lim, and J. C. Ye, "Multi-task vision transformer using low-level chest X-ray feature corpus for COVID-19 diagnosis and severity quantification," *Med. Image Anal.*, vol. 75, p. 102299, 2022.

[49] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[50] Z. Shen, R. Fu, C. Lin, and S. Zheng, "COTR: Convolution in transformer network for end to end polyp detection," in *Int. Conf. on Computer and Communications*, 2021, pp. 1757–1761.

[51] T. S. Mathai, S. Lee, D. C. Elton, T. C. Shen, Y. Peng, Z. Lu, and R. M. Summers, "Lymph node detection in T2 MRI with transformers," in *Medical Imaging 2022: Computer-Aided Diagnosis*, vol. 12033, 2022, pp. 855–859.

[52] H. Li, L. Chen, H. Han, and S. Kevin Zhou, "SATr: Slice attention with transformer for universal lesion detection," in *Med. Image. Comput. Comput. Assist. Interv.*, 2022, pp. 163–174.

[53] F. Shamshad, S. Khan, S. W. Zamir, M. H. Khan, M. Hayat, F. S. Khan, and H. Fu, "Transformers in medical imaging: A survey," *arXiv preprint arXiv:2201.09873*, 2022.

[54] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, and D. Shen, "Transformers in medical image analysis," *Intelligent Medicine*, vol. 3, no. 1, pp. 59–78, 2023.

[55] J. Li, J. Chen, Y. Tang, C. Wang, B. A. Landman, and S. K. Zhou, "Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives," *Med. Image Anal.*, vol. 85, p. 102762, 2023.

[56] R. Byng-Maddick and M. Noursadeghi, "Does tuberculosis threaten our ageing populations?" *BMC Infectious Diseases*, vol. 16, no. 1, pp. 1–5, 2016.

[57] A. L. García-Basteiro, H. S. Schaaf, R. Diel, and G. B. Migliori, "Adolescents and young adults: a neglected population group for tuberculosis surveillance," *European Respiratory Journal*, vol. 51, no. 2, p. 1800176, 2018.

[58] Z. Dong, Q.-Q. Wang, S.-C. Yu, F. Huang, J.-J. Liu, H.-Y. Yao, and Y.-L. Zhao, "Age-period-cohort analysis of pulmonary tubercu-

[59] S. Nhamoyebonde and A. Leslie, "Biological differences between the sexes and susceptibility to tuberculosis," *The Journal of Infectious Diseases*, vol. 209, no. suppl_3, pp. S100–S106, 2014.

[60] K. C. Horton, P. MacPherson, R. M. Houben, R. G. White, and E. L. Corbett, "Sex differences in tuberculosis burden and notifications in low-and middle-income countries: A systematic review and meta-analysis," *PLoS Medicine*, vol. 13, no. 9, p. e1002119, 2016.

[61] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 2117–2125.

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

[63] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Adv. Neural Inform. Process. Syst.*, 2015, pp. 2017–2025.

[64] J. Tang, J. Wang, and J.-F. Hu, "Predicting human poses via recurrent attention network," *Visual Intelligence*, vol. 1, no. 1, pp. 1–9, 2023.

[65] X.-S. Wei, Y.-Y. Xu, C.-L. Zhang, G.-S. Xia, and Y.-X. Peng, "CAT: A coarse-to-fine attention tree for semantic change detection," *Visual Intelligence*, vol. 1, no. 1, p. 3, 2023.

[66] G. Sun, Y. Liu, T. Probst, D. P. Paudel, N. Popovic, and L. V. Gool, "Rethinking global context in crowd counting," *Machine Intelligence Research*, 2023.

[67] Y. Qiu, Y. Liu, L. Zhang, H. Lu, and J. Xu, "Boosting salient object detection with transformer-based asymmetric bilateral U-Net," *IEEE Trans. Circ. Syst. Video Technol.*, 2023.

[68] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Int. Conf. Learn. Represent.*, 2021.

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

[70] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Adv. Neural Inform. Process. Syst.*, 2019, pp. 8024–8035.

[71] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[72] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[73] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.

[74] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Adv. Neural Inform. Process. Syst.*, 2015, pp. 91–99.

[75] M.-M. Cheng, Y. Liu, W.-Y. Lin, Z. Zhang, P. L. Rosin, and P. H. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," *Computational Visual Media*, vol. 5, no. 1, pp. 3–20, 2019.

[76] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learn. Represent.*, 2015.

[77] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[78] Y. Liu, Y.-H. Wu, P. Wen, Y. Shi, Y. Qiu, and M.-M. Cheng, "Leveraging instance-, image-and dataset-level information for weakly supervised instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1415–1428, 2022.

[79] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

losis reported incidence, china, 2006-2020," *Infectious Diseases of Poverty*, vol. 11, no. 04, pp. 62–71, 2022.

Yun Liu received his B.E. and Ph.D. degrees from Nankai University in 2016 and 2020, respectively. Then, he worked with Prof. Luc Van Gool as a postdoctoral scholar at Computer Vision Lab, ETH Zurich, Switzerland. Currently, he is a senior scientist at the Institute for Infocomm Research (I2R), A*STAR, Singapore. His research interests include computer vision and machine learning.

Ming-Ming Cheng received his Ph.D. degree from Tsinghua University in 2012. Then, he did two years research fellow with Prof. Philip Torr in Oxford. He is now a professor at Nankai University, leading the Media Computing Lab. His research interests include computer graphics, computer vision, and image processing. He received research awards, including ACM China Rising Star Award, IBM Global SUR Award, and CCF-Intel Young Faculty Researcher Program. He is on the editorial boards of IEEE TPAMI/TIP.
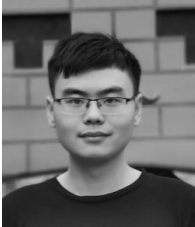
Yu-Huan Wu received his Ph.D. degree from Nankai University in 2022, advised by Prof. Ming-Ming Cheng. He is a scientist at the Institute of High Performance Computing (IHPC), A*STAR, Singapore. He has published 10+ papers on top-tier conferences and journals such as IEEE TPAMI/TIP/CVPR/ICCV. His research interests include computer vision and medical imaging.

Shi-Chen Zhang received his B.E. degree in computer science from Nankai University in 2023. Currently, he is a Ph.D. student in Media Computing Lab, Nankai University, supervised by Prof. Ming-Ming Cheng. His research interests include object detection and semantic segmentation.

Li Liu (Senior Member, IEEE) received her Ph.D. degree from the National University of Defense Technology (NUDT), China, in 2012. She is now a Full Professor with NUDT. Dr. Liu served as a co-chair of many International Workshops along with major venues like CVPR and ICCV. She served as the leading guest editor of the special issues for IEEE TPAMI and IJCV. Her research interests include computer vision, pattern recognition, and machine learning. Her papers currently have over 10,000 citations.

Min Wu (Senior Member, IEEE) received the B.E. degree in computer science from USTC, China, in 2006, and the Ph.D. degree in computer science from NTU, Singapore, in 2011. He is currently a principal scientist with the Institute for Infocomm Research (I2R), A*STAR, Singapore. He received the best paper awards in the IEEE ICIEA 2022, the IEEE SmartCity 2022, the InCoB 2016, and the DASFAA 2015. He also won the CVPR UG2+ challenge in 2021 and the IJCAI competition on repeated buyers prediction in 2015. His current research interests include machine learning, data mining, and bioinformatics.