Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-14790-7

# A cross-modal deep metric learning model for disease diagnosis based on chest x-ray images

Yufei Jin[1,2] · Huijuan Lu[1,2] · Zhao Li[3] · Yanbin Wang[4]

Received: 11 April 2022 / Revised: 12 August 2022 / Accepted: 5 February 2023

**Abstract**
The emergence of unknown diseases is often with few or no samples available. Zero-shot learning and few-shot learning have promising applications in medical image analysis. In this paper, we propose a Cross-Modal Deep Metric Learning Generalized Zero-Shot Learning (CM-DML-GZSL) model. The proposed network consists of a visual feature extractor, a fixed semantic feature extractor, and a deep regression module. The network belongs to a two-stream network for multiple modalities. In a multi-label setting, each sample contains a small number of positive labels and a large number of negative labels on average. This positive-negative imbalance dominates the optimization procedure and may prevent the establishment of an effective correspondence between visual features and semantic vectors during training, resulting in a low degree of accuracy. A novel weighted focused Euclidean distance metric loss is introduced in this regard. This loss not only can dynamically increase the weight of hard samples and decrease the weight of simple samples, but it can also promote the connection between samples and semantic vectors corresponding to their positive labels, which helps mitigate bias in predicting unseen classes in the generalized zero-shot learning setting. The weighted focused Euclidean distance metric loss function can dynamically adjust sample weights, enabling zero-shot multi-label learning for chest X-ray diagnosis, as experimental results on large publicly available datasets demonstrate.

**Keywords** Chest X-ray · Generalized zero-shot learning · Deep metric learning · Cross-modal · Multi-label classification

## 1 Introduction

Multi-label zero-sample learning is a challenging problem. Popular similarity-based methods appear incapable of solving the multi-label problem, despite their enormous success with single-label tasks. Our goal is to identify all labels in a chest X-ray image, including unseen classes. Chest X-ray images are used to detect pneumonia, but this requires the expertise of radiologists or other specialists. And every day, physicians must read a large

✉ Huijuan Lu
hjlu@cjlu.edu.cn

Extended author information available on the last page of the article.

Published online: 15 March 2023

number of chest X-rays, which can cause visual fatigue. It will influence the doctor's diagnosis to some degree. In this paper, the significance of generalized zero-shot learning is that there are no learnable samples for new and rare diseases. The general deep learning models will make erroneous judgments based on learned image knowledge when they encounter classes they have not previously encountered. In some instances, one disease may contribute to the development of another disease or its symptoms. The purpose of this paper is to accurately diagnose each chest disease through chest X-ray images, which is also of great significance for the control and diagnosis of new coronary pneumonia.

With the continued advancement of deep learning and neural networks in recent years, the application of convolutional neural networks in image detection, gene classification and medical imaging has produced astounding results. Hinton et al. [21] introduced the concept of neural network in 2006 and defined its essence as supervised learning. As an important branch of artificial intelligence, deep learning has opened up an emerging research direction in the field of modern computers. Salim et al. [47] adopted an image annotation and segmentation model to annotate and segment hands from radiological images, first removing the background, then learning significant features via a regression neural network architecture, and finally producing bone age predictions. Sharma et al. [49] used different deep convolutional neural networks to extract different structures of features from chest X-ray images and classify the images to detect whether patients have pneumonia. Using deep convolutional neural networks, data augmentation, and multi-task learning, Gabruseva et al. [12] implemented a computational method for pneumonia region detection, which achieved one of the best results in the North American pneumonia detection challenge.

Complex problems can be effectively mastered using deep learning. However, one of the drawbacks of deep learning is that it requires a large of labeled training data, and the model focuses on the samples that have already appeared in the labeled category during the training. In many real-world application scenarios, however, there are only a few numbers of data samples, and the model is even required to classify unseen instance categories. The problem of sparse data samples can be solved by few-shot learning methods. Cheng et al. [8] proposed a Task-wise Attention (TWA) module and a Part Complementary Learning (PCL) module for the few-shot classification task in order to specialize the base learner to respond to different inputs for different tasks. Besides classification tasks, few-shot segmentation tasks have also been developed extensively. Lang et al. [35] added a second base learner to the conventional Few-shot Segmentation (FSS) model, allowing the coarse results of the parallel outputs of the two learners to be adaptively integrated to produce accurate segmentation forecast. Correspondingly, if there is only one labeled sample, it is called One-shot Learning. If the sample labeling learning is not performed for this class, it is Zero-shot Learning. Inspiring zero-shot learning is the human learning strategy of inferring others. The issue and concept of zero-shot learning was proposed by Lampert et al. [33]. Given its ability to identify novel classes, ZSL may have potential applications in radiodiagnosis, particularly for diagnosing rare diseases from radiographic images that lack imaging information and contain only text descriptions. Zero-shot learning is a research hotspot at present, but it also faces many challenges. During the training of a zero-shot learning model, the known classes have visual image features while the unknown classes only have semantic information. However, during testing, the model needs to identify visual features from unknown classes. Enhancing the feature extraction ability of the model can also improve the performance. Gu et al. [15] proposed a novel deep dual-channel neural network (DCNN), which utilized the dual-channel network to extract more comprehensive image features and achieved improved performance. Gu et al. [17] used information richness measures to extract PM2.5-related high-level features. Gu et al. [16] investigated the feasibility of ensemble meta-learning to improve the extraction of key features from deep neural networks through fine-tuning.

Transferring knowledge across domains is one of the keys to zero-shot learning. Currently, more research focuses on samples with a single label; for samples with multiple labels, the model must be more complex, and there are fewer studies. However, chest diseases, such as pneumonia, frequently result in complications and severe consequences if they are not correctly diagnosed.

To overcome the above challenges, we propose a cross-modal deep metric learning method to solve the disease classification problem on multi-label chest X-ray images. Cheng et al. [7] demonstrated that metric learning can facilitate models to be more discriminative, making images from the same class are mapped closely to each other and the images of different classes are mapped as farther apart as possible. Our model is comprised of a visual encoder, a semantic encoder, and a deep regression module that learns distance metrics between cross-modal features by employing labeled word vectors as prototypes for each class. We use a large public pneumonia dataset to evaluate our model. The main contributions of this paper are as follows:

- The proposed network is a two-stream network for multiple modalities, which projects visual embeddings and semantic vectors into a common space without requiring offline training of the visual feature encoder.
- For the construction of the learning model, a loss function based on the weighted focused Euclidean distance metric is suggested. This loss function is capable of dynamically adjusting sample weights and promoting the connection between samples and semantic vectors corresponding to their positive labels, thereby improving accuracy.
- In the same configuration of seen and unseen classes, the AUROC value of the CM-DML-GZSL model exceeds 0.80 for the seen class and improves by 2 percent for the unseen class compared to the suboptimal model.

## 2 Related work

### 2.1 Multi-label learning

Single-label learning is a special case of multi-label learning, in which each sample has a single label [26]. Traditional supervised learning is primarily single-label learning, whereas real-world target samples are frequently more complex. Therefore, single-label recognition can no longer satisfy the current practical applications. At present, multi-label learning has been studied extensively in image classification, object detection, video classification, and gene function prediction [9, 30, 45, 60]. The objective of this paper is to improve the classification accuracy of multi-label chest X-ray images.

The exponentially complex nature of the label space is one of the primary obstacles of multi-label learning. In order to cope with the explosive growth of the output space, the existing solutions are primarily divided into two ways [64]. One approach is to convert the multi-label problem into other learning scenarios, such as decomposing the multi-label problem into multiple independent binary classification problems [3] or sorting related tags and irrelevant tags [11]. Boutell et al. [3] presented a framework for addressing the problem of multiple labels. The basic idea of their paper is to decompose the multi-label learning problem into $q$ independent binary classification problems, where each binary classification problem corresponds to a label in the label space. And they implemented new criteria for assessing individual examples, class recall and precision, and overall accuracy. Due to the fact that the label ranking method is severely constrained by the calibration scale, the author of [11] introduces a calibration label that represents the boundary between relevant and irrelevant labels. The alternative is to process multi-label data directly using algorithms [5, 6]. Chen et al. [6] utilized the attention to extract different feature vectors from distinct label objects in the image, and then used Graph Neural Network (GNN) to perform in-depth processing of pixel feature vectors. Finally, the output of feature vectors to the classifier for prediction. Chen et al. [5] proposed a new multi-label Graph Convolutional Network (ML-GCN) with the innovation of learning interdependent object classifiers from label features via Graph Convolutional Network (GCN).

### 2.2 Zero-shot learning

Zero-Shot Learning (ZSL) seeks to identify instances from unknown or unseen categories without a single training example. This is a difficult task. In conventional zero-shot learning, there is no overlap between the training set and the test set. However, in practical applications, zero-shot learning tasks are more generalized. The training set contains only categories that have been observed, whereas the test set includes both observed and unseen categories. Zero-shot learning is a special scenario of transfer learning, which requires knowledge transfer between seen and unseen categories to complete the learning. Appropriate auxiliary information can facilitate the transfer of knowledge. Auxiliary information may consist of semantic attributes [34], a visual description [2] or word vectors [43]. Not only classification tasks [42, 58, 59], video object segmentation also has important research significance in the field of zero-shot learning. Wang et al. [55] proposed an attentive graph neural network (AGNN) for zero-shot video object segmentation (ZVOS), recasting ZVOS as a message-passing-based end-to-end graph information fusion method. Lu et al. [40] developed the CO-attention Siamese Network (COSNet) to address the task of unsupervised video object segmentation from a holistic perspective. Lu et al. [41] is an extension of [40] that generalizes the originally proposed pair-wise attention to a group-wise format to significantly enhance the algorithm and make more efficient use of global information.

In particular, the current mainstream zero-shot learning methods mainly adhere to the following three concepts: i) Learning the mapping relationship between visual space and semantic space [46, 63]. The classic ALE model [1] proposes a label embedding framework to solve the problem of category prediction directly. Inspired by the label embedding method, Akata et al. [2] proposed a Structured Joint Embedding (SJE) framework and the SJE-based nonlinear model LatEm [56]. ii) Acquiring an understanding of the mapping relationship between semantic and visual space [31, 63, 65]. Shigeto et al. [51] first proposed and demonstrated that mapping labels to visual space is effective for preventing hubness problems. iii) Learning the mapping of visual and semantic spaces to a hidden space shared by both [62]. In contrast to the previous two methods, zero-shot learning involves projecting both visual features and semantic coding into the public space, and then obtaining the final recognition result via distance measurement.

The objective of Generative Adversarial Network (GAN) techniques is to generate uncommonly positive samples. By synthesizing chest X-ray images of COVID-19-positive cases. Zunair et al. [68] addressed the severe lack of chest radiographic data in COVID-19-positive patients. To solve the class imbalance problem, conditional images of unpaired images were used to generate synthetic dermoscopic images in [67]. Recent research has demonstrated that GANs facilitate the study of zero-shot learning. It can generate visual characteristics of invisible classes using semantic attributes and achieves the objective of zero-shot learning, which is to differentiate between visible and invisible categories. The work of Xian et al. [57] proposed using semantic information and the GAN framework to generate CNN features of unseen categories. Then, they used Wasserstein GAN and classification loss to generate CNN features for training the classifier that were discriminative. Sariyildiz et al. [48] used a generation model to generate both seen and unseen category samples, and used the generated samples and seen categories as training set to train classifier. The zero-shot problem is converted into a supervised learning problem. Because Generative Adversarial Network training is not sufficiently stable, adding additional regular terms may lead to unpredictable training results. Shen et al. [50] proposed a zero-shot learning method based on Generative Flow model for the first time, which can directly describe the probabilistic characteristics of sample space and has a relatively simple and stable training process. Although image zero-shot learning has achieved great progress, the majority of current research focuses on zero-shot single-label classification. The application of the zero-shot concept to multi-label learning remains an open question [26].

### 2.3 Multi-label zero-shot learning

Current research on multi-label zero-shot image classification is quite limited, which is more difficult than its single-label equivalent. Fu et al. [10] enumerated all possible multi-label combinations in the semantic space and transformed the multi-label zero-shot classification problem into a single-label zero-shot classification problem in order to simplify multi-label processing. This method's greatest limitation is that the number of labels should not be excessive. When the number of sample categories in the data set increases, the number of combined labels will increase exponentially, causing the output space to grow explosively. Gaure et al. [13] proposed a probability model, which learning label attributes directly using the label co-occurrence statistics and predicting invisible classes during testing using label co-occurrence. In recent years, Lee et al. [36] proposed a framework that combines knowledge graph to describe the relationship between multiple labels. This model learns a knowledge propagation mechanism from semantic label space to describe the interdependence mechanism between seen categories and unseen categories. The method of knowledge graph is also adopted. Huang et al. [23] constructed a knowledge graph by combining the target dataset and ImageNet dataset, and designed a Relational Graph Convolutional Network (RGCN) to achieve information dissemination and knowledge transfer between categories. Gupta et al. [18] were the first to solve the problem of multi-label feature synthesis in a generalized zero-shot setting by utilizing Generative Adversarial Networks in recent research. This paper is based on deep metric learning to solve the problem of multi-label zero-shot classification. Metric learning is more and more widely used in image classification, few-shot learning, zero-shot learning as well as other fields, with excellent results [4, 28, 29].

### 2.4 Deep metric learning

Recent research has demonstrated that the powerful generalization performance of distance metric learning is applicable to multi-label learning. In particular, if the labels for a given example are similar in the label space, they should also be closer in the embedded feature space. For instance, Gouk et al. [14] trained an embedder that can transform instances into a feature space employing the square Euclidean distance. Similarly, Zhang et al. [66] proposed a novel multi-view distance metric learning approach to dealing with the multi-label image classification issue. To preserve the inherent geometric structure of the data in the low-dimensional feature space, they introduced a manifold regularization with the adjacency graph being constructed based on all labels.

The purpose of metric learning is to learn the transformation from image space to feature embedding space. Feature vectors of the same category are closer in the metric space, and feature vectors of different categories are farther apart in the metric space. Class-Prototype Discriminative Network (CPDN) was proposed by Huang et al. [24] for generalized zero-shot learning. They generated class prototype for each category and distinguished class prototype through metric network, and obtained the final classification result according to the similarity between samples and class prototype. Using two triplet networks, Ji et al. [27] constructed a new deep metric learning framework. Firstly, semantic information is mapped to the visual space, and then a Dual-Triplet Network (DTNet) is used to learn visual-semantic mapping in order to guarantee the complete discovery and utilization of data information. In the multi-label image classification problem, Li et al. [38] constructed a unified metric space of potential image and label, making the distance between image and corresponding target label smaller than the distance between image and other adjacent labels, as well as the distance between other images corresponding to these adjacent labels. Metric learning methods are based on learning a mapping function, the quality of the mapping function directly determines the effectiveness of metric learning methods. The simple and efficient efficiency of deep metric learning has been proved [39, 61], but it has rarely been used in multi-label zero-shot scenarios. Metric learning can effectively solve multi-label problems, and future research will focus more on this topic.

## 3 Method

Deep learning requires a large number of samples, but in certain cases, such as the emergence of new diseases, the number of samples that can be obtained is extremely limited or even no samples. In this instance, this paper proposes a new method to solve the zero-shot learning challenge using cross-modal deep metric learning.

### 3.1 Problem definition

Suppose that the training set (seen classes) $\mathcal{S} = \{(x_i^S, y_i^S), i = 1, 2, ..., N_S\}$ is given, where $x_i^S \in \chi^S$ is the $i$-th data samples with a corresponding seen class label as $y_i^S$, here $y_i^S = \{0, 1\}^S$ is the corresponding multi-hot labels from the set of $\mathcal{S}$ seen class labels $C^S$. $N_S$ represents the number of samples of all seen classes. $C^S$ represents the number of categories of seen classes. In traditional zero-shot learning, given the testing set $\mathcal{U} = \{(x_i^U, y_i^U), i = 1, 2, ..., N_U\}$ from the unseen classes, where $x_i^U \in \chi^U$ and $y_i^U = \{0, 1\}^U$ is the $i$-th unseen sample and its label, respectively. Moreover, we use BioBERT [37] vectors of the class names as the attribute embeddings $a(y)$, as in [19]. BioBERT is a domain-specific language representation model pre-trained on large biomedical corpora. After pre-training on biomedical corpora, BioBERT has achieved excellent results in many biomedical text mining tasks. Different from the traditional ZSL, the purpose of GZSL is to learning a mapping function $f(\boldsymbol{x}_i) : \chi \rightarrow \{0, 1\}^S$ aided by the attribute embedding $\{\boldsymbol{A}_i\}_{i=1}^C$. Such mapping function can be adjusted at test time to accommodate $\mathcal{U}$ unseen classes (with embedding $\boldsymbol{A}_i^U \in \mathbb{R}^{C^U \times d_a}$). That is $f(\boldsymbol{x}_i) : \chi \rightarrow \{0, 1\}^U$ for ZSL and $f(\boldsymbol{x}_i) : \chi \rightarrow \{0, 1\}^C$ for GZSL setting. Here, $C = C^S + C^U$ represents the total number of seen and unseen classes.

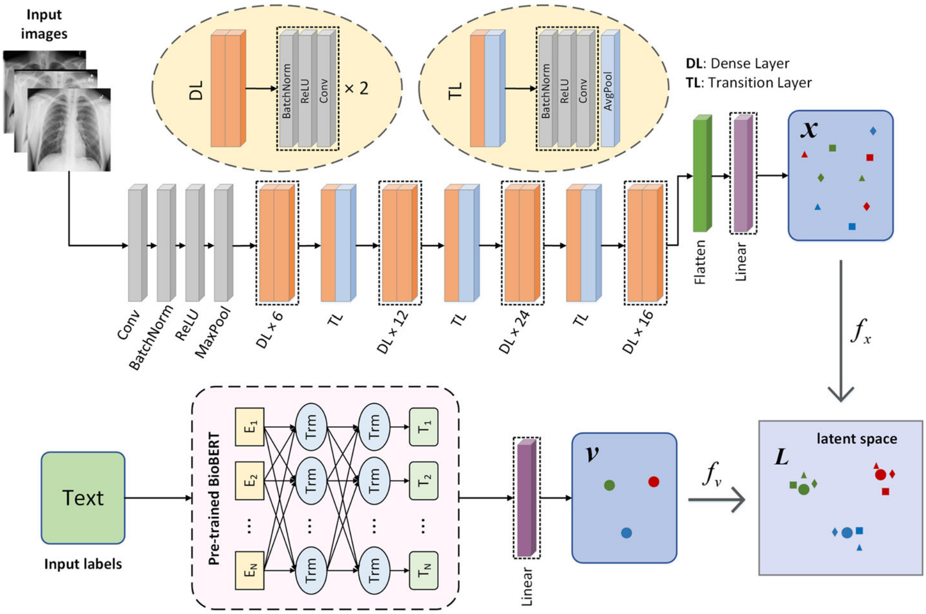


**Fig. 1** Overview of the CM-DML-GZSL model. It includes a trainable visual encoder DenseNet121 for learning visual representations of chest X-rays, and a semantic space. For the input image and its corresponding label, the label semantic word vector is extracted by the BioBert network. The visual vector and the tag word vector are respectively projected to the same latent space through the mapping module. Finally, the entire model is put through end-to-end training

### 3.2 Network architecture

An overview of the network structure is shown in Fig. 1. It consists of a visual encoder that can be trained, a semantic encoder that is fixed, and a mapping module. Next, we will describe each component.

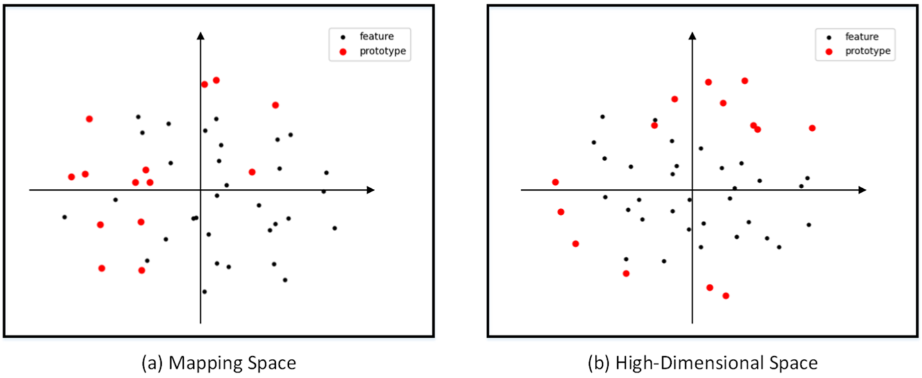
Visual Encoder: DenseNet121 [22] is selected as the image feature extractor, which consists of 4 Dense Blocks with a convolutional layer and an average pooling layer between each two Dense Blocks. We replace the final classification layer as a feature extraction layer, with the extracted features serving as the visual vector $f(\boldsymbol{x})$ of the input image.

Semantic Encoder: The corresponding label $y$ of the image is extracted from the pre-trained BioBert network. During model training, the BioBert network no longer updates parameters.

Mapping Module: It is used to align the vectors of two different modes, and project the vectors of two different forms from high-dimensional to low-dimensional common latent space. The specifics are elaborated upon in the following section.

### 3.3 Mapping module

The semantic and visual vectors originate from two distinct sources. In order to bridge the gap between these two patterns, the model first obtains the projection between high-dimensional visual space $X$ and semantic space $V$ via an embedding-based method. To



(a) Mapping Space

(b) High-Dimensional Space

**Fig. 2** Mapping space (a) and high-dimensional space (b) feature visualization

establish the classification model, the mapping between the visual feature space and the semantic space must be established, but the Hubness problem is inherent to high-dimensional spaces. Therefore, we mapped the image features and tag word vectors to an additional low-dimensional common embedding space through two mapping functions. As shown in Fig. 2, the left map is unquestionably superior for classification when visualizing the features of the mapping space and the high-dimensional space.

These two mapping functions are neural network-learned models of deep regression. As shown in Fig. 3, each model consists of three FC layers with dimensions of 512, 256, and 100. Each layer consists of a fully connected (FC) layer, and there is a Leaky Rectified Linear Unit (Leaky ReLU) function between every two layers. Particularly, FC layer can assist with the migration of model representation capabilities.

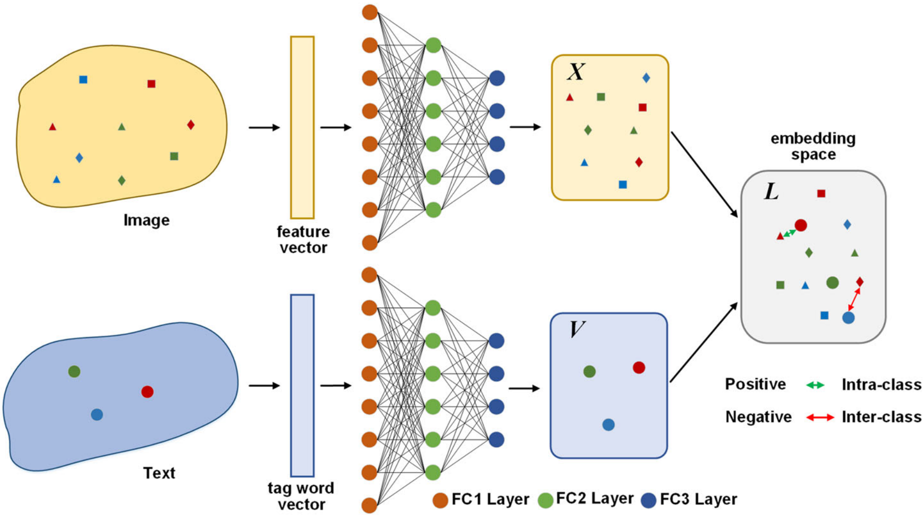


**Fig. 3** The architecture of the mapping module. The visual vector in the visual space $X$ and the tag word vector in the semantic space $V$ are respectively mapped into the common space through the projection network

### 3.4 Loss function

The objective of the model is to create a metric space that facilitates and improves the classification task, such that the distance between relevant labels of a query image is less than that between irrelevant labels. During the training process, the training set of the model comes from the input image $x^S$ of the visible class $\mathcal{S}$ and the corresponding label $y^S$. $y_j^S = \{0, 1\}_{j=1}^{S}$, 0 means that the sample does $\notin$ class $j$, and 1 means that the sample $\in$ class $j$. The auxiliary information semantic vector is used as the prototype of each class to calculate the distance from each sample. Since this is a multi-label classification task. First, we choose the weighted Euclidean distance as the basic distance evaluation metric, as shown in (1), (2) and (3).

$$dist(\boldsymbol{x}, \boldsymbol{y}') = \sqrt{\sum_{i=1}^{N} \frac{\mathbf{w}_i}{W} (\boldsymbol{x}_i - \boldsymbol{y}'_i)^2} \tag{1}$$

$$W = \sum_{i=1}^{N} \mathbf{w}_i \tag{2}$$

$$\mathbf{w}_i = \frac{r}{2} |\boldsymbol{x}_i - \boldsymbol{y}'_i|^2 \tag{3}$$

In the above Equations, $\boldsymbol{x}$ represents the visual feature vector, $\boldsymbol{y}'$ represents the prototype feature, $\frac{\mathbf{w}_i}{W}$ is the weight of different coordinates, here $r = 1$. In addition to the weighted Euclidean distance, we also introduce the following commonly used distance evaluation metrics for discussion.

Manhattan Distance: It is used to indicate the sum of the absolute wheelbases of two points on the standard coordinate system. The Manhattan distance between the visual feature vector $\boldsymbol{x}$ and the prototype feature $\boldsymbol{y}'$ is:

$$dist_M(\boldsymbol{x}, \boldsymbol{y}') = \sum_{i=1}^{N} |\boldsymbol{x}_i - \boldsymbol{y}'_i| \tag{4}$$

Chebyshev Distance: As a measure in vector space, the largest value of the difference between two points' respective coordinates is what is referred to as the distance between two points in vector space. The Chebyshev distance between $\boldsymbol{x}$ and $\boldsymbol{y}'$ is:

$$dist_C(\boldsymbol{x}, \boldsymbol{y}') = \max_i (|\boldsymbol{x}_i - \boldsymbol{y}'_i|) \tag{5}$$

It can also be expressed as:

$$dist_C(\boldsymbol{x}, \boldsymbol{y}') = \lim_{k \to \infty} \left( \sum_{i=1}^{N} |\boldsymbol{x}_i - \boldsymbol{y}'_i|^k \right)^{\frac{1}{k}} \tag{6}$$

Lance and Williams Distance: It is considered a weighted version of Manhattan distance. Utilizing the visual feature vector $\boldsymbol{x}$ and the prototype features $\boldsymbol{y}'$ an example, the Lance and Williams distance between them is:

$$dist_L(\boldsymbol{x}, \boldsymbol{y}') = \frac{1}{N} \sum_{i=1}^{N} \frac{|\boldsymbol{x}_i - \boldsymbol{y}'_i|}{|\boldsymbol{x}_i| + |\boldsymbol{y}'_i|} \tag{7}$$

Secondly, due to the unbalanced number of samples, we combine the idea of Focal Loss to reduce the weight of simple negative samples and mine difficult samples. The Focal Loss formula is shown in (8):

$$FL(p_t) = -(1 - p_t)^{\alpha} \log(p_t) \tag{8}$$

Where, $(1 - p_t)^{\alpha}$ is the modulating factor, $\alpha$ is the focusing parameter and $\alpha \geq 0$.

Finally, assuming that $d_{y_p}$ and $d_{y_n}$ are the distances between the positive label and the negative label and the prototype $a(y)$, respectively, the following metric relationship needs to be satisfied:

$$L_{Metric} = \frac{1}{N} \sum_{\forall x \in \chi^S} \frac{1}{S} \sum_{y_p \in Y_p} \sum_{y_n \in Y_n} \max(D, 0) \tag{9}$$

$$D = \delta - \beta E + \left( d_{y_p}^{\alpha} \log_{10} e^{d_{yp}} - d_{y_n}^{\alpha} \log_{10} e^{d_{yn}} \right) \tag{10}$$

Where, $N$ is the number of all samples, $S$ is the total number of visible classes, $D$ represents the distance relationship between relevant labels and irrelevant labels, ideally, we wish to meet two conditions. First, $d_{y_n}$ should be greater than $d_{y_p}$, and second, the difference between them should be at least a margin value. In addition, $d$ is the weighted Euclidean distance, $\delta$ is the margin value, which decreases with the continuous training of the model, $\beta$ is the weight of $\delta$ reduction is 0.03, $E = \lfloor epoch/10 \rfloor$, $\alpha$ is the focusing parameter, which is set to 2 here. If the distance of the irrelevant labels is greater than the margin value plus the distance of the relevant labels, the loss is 0, otherwise the loss is $D$. That is, $d_{y_p}$ is at least a margin value smaller than $d_{y_n}$.

## 4 Experimental results and analysis

In this section, numerous experiments are conducted to demonstrate that our proposed model is effective for multi-label zero-shot classification. The experimental dataset and experimental parameters are presented first, followed by the main results of the contrasting experiments and ablation experiments.

### 4.1 Dataset and experimental setup

The experimental setup and data set are described in detail below. We used the Chest X-ray14 dataset provided by the NIH Research Institute [44], which is a multi-label data set and there is a chest X-ray image containing multiple diseases. At the same time, it is also the largest and most disease-type chest X-ray image dataset.

#### 4.1.1 Dataset

The ChestX-ray14 dataset consists of 112,120 frontal X-ray images from across 30,805 patients, which is an extension of the 8 common disease dataset used in the paper by Wang et al. [54]. Among them, the resolution of each image is 1024 × 1024. In this paper, we do not consider normal samples (healthy samples, excluding any disease) and ensure that each image is associated with 14 possible classes. We split the dataset randomly to obtain a training set (70%), validation set (10%), and test set (20%). Then, we divide the seen (10 in total) and unseen (4 in total) classes. Specifically, the seen classes are Atelectasis, Effusion, Infiltration, Mass, Nodule, Pneumothorax, Consolidation, Cardiomegaly, Pleural Thickening, and Hernia. The rest are unseen classes, they are Edema, Pneumonia, Emphysema, and

**Table 1** Experimental setup

| Parameter name | Parameter value |
|---|---|
| pre-training config | DenseNet121 |
| optimizer | Adam |
| base learning rate | 1e-4 |
| weight decay | 3e-4 |
| optimizer momentum | $\beta_1, \beta_2 = 0.9{,}0.999$ |
| batch size | 32 |
| training epochs | 40 |

Fibrosis. According to the inductive zero-shot setting, we ensure that the samples in the training set do not have unseen class label images. The final training set contained 30,758 images, the validation set contained 4,474 images, and the test set contained 10,510 images. It is consistent with the setting in [19].

#### 4.1.2 Experimental setup

All experiments used the open-source Pytorch neural network framework, and the neural networks are trained on an Intel Core i7-10700 processor and an NVIDIA GTX3060 GPU accelerator. We utilized data augmentation to enhance network precision and avoided overfitting by expanding the size of the training dataset, including random cropping and horizontal flipping. The initial learning rate is set to 0.0001, and every 10 epochs, the initial learning rate will be reduced by a factor of 0.1, and the batch size is 32, as shown in Table 1. Additionally, we trained the network for a total of 40 epochs using the Adam optimizer during training, which took approximately 7 hours. Our proposed model is trained in an end-to-end manner from the beginning without pre-training.

For feature extraction, we used a pretrained DenseNet121 convolutional neural network with an input image size of 224×224×3. First, we resized the input image to 256×256 and then cropped it to 224×224, we removed the final classification layer of the pretrained DenseNet121 and used the resulting network as a visual encoder, generating visual features $\boldsymbol{x} \in \mathbb{R}^{1024}$. Then high-dimensional features need to be mapped into low-dimensional space, and the visual mapping module is parameterized as a three-layer feedforward neural network: $f(x) \longrightarrow \text{fc1} \longrightarrow \text{LeakyReLU} \longrightarrow \text{fc2} \longrightarrow \text{LeakyReLU} \longrightarrow \text{fc3} \longrightarrow \boldsymbol{x}, \boldsymbol{x} \in \mathbb{R}^{100}$. Similarly, semantic vectors are also projected into a potentially low-dimensional common space.

### 4.2 Results and analysis

#### 4.2.1 Main results

In this part, we will study the effects of parameter $\alpha$ and parameter $\beta$ on the performance of chest disease classification based on cross-modal deep metric learning. Then, four distinct zero-shot multi-label chest X-ray image disease detection techniques were compared. We evaluated the comparison methods using the Area Under the Receiver Operating Characteristic Curve (AUROC) score. AUROC is appropriate for unbalanced samples and is a widely utilized performance evaluation standard for multi-label classification. It is a number between 0 and 1. The more closely the AUROC value approaches 1, the more accurately the classifier can distinguish between positive and negative samples. The Harmonic Mean (HM) is another performance metric that measures the inherent bias of GZSL-based methods with respect to the seen class; if the GZSL method is biased toward the observed class, the HM score decreases. The Equation is shown in (11):

$$HM = 2 \times \frac{AUC_s \times AUC_u}{AUC_s + AUC_u} \tag{11}$$

First, we discussed the value of $\alpha$. $\alpha$ as a focusing parameter, the weight of loss can be controlled by the modulating factor, so that the weight of hard samples (low score) tends to 1, while the weight of simple samples (high score) will be scaled to a small value. When the loss weight of hard samples remains unchanged, and the loss weight of simple samples decreases, it is equivalent to increasing the weight of hard samples. Figure 4 shows the effect of parameter $\alpha$ on the performance of chest disease classification based on cross-modal deep metric learning. When the parameter $\alpha = 0$, it is the ordinary distance metric loss function. As $\alpha$ increases, the performance of the model also improves. When $\alpha = 2$, both the seen and unseen class achieve an optimum. When the parameter $\alpha$ is greater than 2, the AUROC of seen and unseen classes show a decreasing trend, especially for unseen classes. The value of $\alpha$ for subsequent experiments was set to 2.

The goal of metric learning is to make the distance between the same classes closer and the distance between different classes farther. With the continuous training of the model, the distance between the same class is decreasing. Since the experimental data set is multi-label samples, appropriately reducing the distance threshold can make the model converge faster. The $\beta$ value is the decreasing weight of the threshold $\delta$, which makes the threshold constantly decrease. Figure 5 shows the effect of parameter $\beta$ on the performance of chest disease classification based on cross-modal deep metric learning. It can be seen that when $\beta$ is less than 0.02, AUROC of seen class and unseen class is in an upward trend. When $\beta$ is greater than 0.03, the AUROC of seen and unseen classes decreases, especially the unseen classes. This shows that a small or large $\beta$ value is not conducive to model training. It can

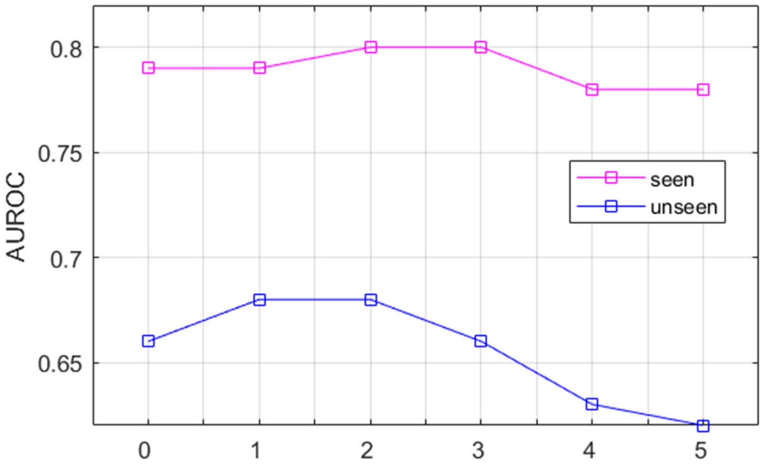


**Fig. 4** Line chart of relationship between parameter $\alpha$ and zero-shot multi-label chest disease classification performance. The abscissa is the value of parameter $\alpha$, and the ordinate is the value of AUROC. The purple line represents seen classes and the blue line represents unseen classes. $\beta$ is fixed at 0.03 during the experiment

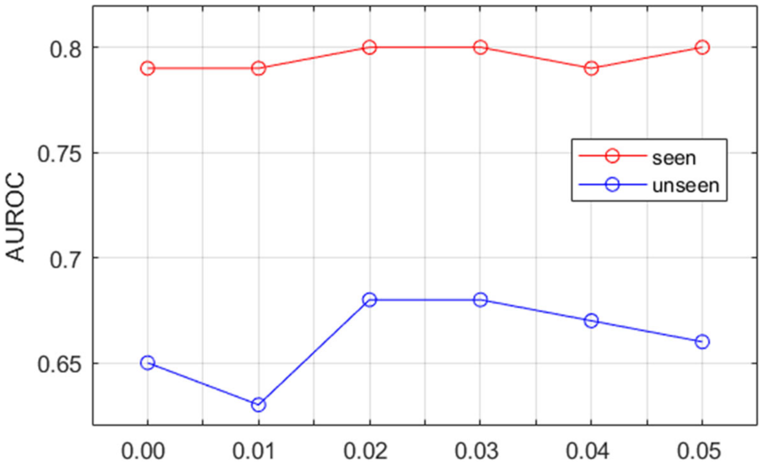


**Fig. 5** Line chart of relationship between parameter $\beta$ and zero-shot multi-label chest disease classification performance. The abscissa is the value of parameter $\beta$, and the ordinate is the value of AUROC. The red line represents seen classes and the blue line represents unseen classes. $\alpha$ is fixed at 2 during the experiment

be seen from the broken line that the parameter 0.03 adopted in this paper is a relatively great equilibrium point. The $\beta$ value for subsequent experiments was set to 0.03.

In this experiment, we compared the proposed CM-DML-GZSL model with some of the most commonly employed and excellent zero-shot learning models. These include LFSA, MLZSL, LEA-ZSL, DEM-ZSL and CXR-ML-GZSL methods, and the specific models and experimental results are shown in Table 2. We compared all current cutting-edge multi-label generalized zero-shot learning methods. LESA trains a set of label-independent shared attention, and then passes a compatibility function to distinguish labels according to the chosen attention. MLZSL learns an information propagation mechanism from the semantic label space for modeling the interdependencies between seen and unseen class labels. LEA-ZSL embeds visual vectors into semantic vectors and uses distance functions for classification. DEM-ZSL embeds the semantic vector within the visual vector, and bases classification on the distance between the test sample and the class semantic vector. CXR-ML-GZSL ranks by the correlation scores of image feature vectors and semantic vectors. Convolutional neural network models are usually used to extract features, and here DenseNet121 is used as the visual feature extractor and BioBERT as the semantic encoder. From the results, using the weighted focused Euclidean distance loss function is discriminative and general for both seen and unseen classes on the chest X-ray image dataset. Compared with other methods, our method has significant gains on many individual classes. This is advantageous due to the co-mapping of visual vectors and semantic vectors into a common low-dimensional space, avoiding inherent and hubness problems in high-dimensional space, and the adaptive loss function promotes the connection between samples and their corresponding semantic vectors of positive labels. We improved the average AUROC gain for both seen and unseen classes in a more straightforward manner, demonstrating the efficacy of our model. This also validated our belief that selecting an appropriate embedding space is the key to understanding the function of deep embedding models and that utilizing distance metric functions is an effective strategy for zero-shot learning.

**Table 2** AUROC class comparison for all disease classes in the test set

| Method | Seen mean | Unseen mean | Atelectasis | Cardiomegaly | Effusion |
|---|---|---|---|---|---|
| LESA [25] | 0.51 | 0.52 | 0.54 | 0.51 | 0.48 |
| MLZSL [36] | 0.72 | 0.54 | 0.70 | 0.76 | 0.76 |
| LEA-ZSL [1] | 0.78 | 0.59 | 0.76 | 0.89 | 0.81 |
| DEM-ZSL [65] | 0.79 | 0.64 | 0.75 | **0.91** | 0.82 |
| CXR-ML-GZSL [19] | 0.79 | 0.66 | 0.76 | 0.90 | **0.83** |
| OURs | **0.80** | **0.68** | **0.77** | **0.91** | **0.83** |

| Method | Consolidation | Nodule | Pleural thickening | Mass | Pneumothrax |
|---|---|---|---|---|---|
| LESA [25] | 0.54 | 0.50 | 0.52 | 0.52 | 0.51 |
| MLZSL [36] | 0.68 | 0.69 | 0.66 | 0.67 | 0.76 |
| LEA-ZSL [1] | **0.69** | 0.72 | 0.71 | 0.76 | 0.83 |
| DEM-ZSL [65] | 0.68 | 0.75 | 0.72 | 0.77 | 0.83 |
| CXR-ML-GZSL [19] | **0.69** | 0.75 | 0.72 | 0.80 | 0.83 |
| OURs | **0.69** | **0.77** | **0.74** | **0.81** | **0.84** |

| Method | Infiltration | Hernia | *Pneumonia* | *Edema* | *Emphysema* | *Fibrosis* |
|---|---|---|---|---|---|---|
| LESA [25] | 0.50 | 0.47 | 0.50 | 0.54 | 0.50 | 0.53 |
| MLZSL [36] | 0.64 | 0.90 | 0.52 | 0.51 | 0.58 | 0.54 |
| LEA-ZSL [1] | 0.69 | 0.92 | 0.60 | **0.68** | 0.68 | 0.41 |
| DEM-ZSL [65] | 0.68 | 0.94 | **0.64** | **0.68** | 0.72 | 0.53 |
| CXR-ML-GZSL [19] | **0.70** | 0.90 | 0.62 | 0.67 | 0.74 | 0.60 |
| OURs | **0.70** | **0.96** | 0.62 | 0.66 | **0.79** | **0.67** |

In this table, we showed the performance AUROC for all seen and unseen classes. Our method has significant gains on many individual classes. Compared to other methods, our method achieves up to 16% AUROC gain on average performance on unseen classes and 2% AUROC gain over the second best method (CXR-ML-GZSL). Italicized classes refer to unseen classes, and bold numbers represent the best AUROC in the corresponding column

We conduct comparative experiments on different distance evaluation metrics, as shown in Table 3. It can be seen from the experimental results that Euclidean distance is the most widely employed and optimal distance measurement method. In comparison to the Manhattan distance, Chebyshev distance, and Lance and Williams distance, Euclidean distance achieves the best classification results in this experiment. However, it treats the effects of different dimensions of the sample equally, especially for the unseen class samples with missing samples, which are easily influenced by the features of the seen class samples. Therefore, the weighted Euclidean distance can more accurately reflect the sample's overall characteristics.

Figure 6 is an example prediction for 12 chest X-ray images in the test set. According to the analysis results of the dataset, the number of samples with more than 3 (excluding 3) true labels is few, so we select the top three predicted values on each test image. It can be seen from the Fig. 6. that our model can still detect unseen classes even when the number of

**Table 3** Results of different distance evaluation metrics

| Distance function | Seen | Unseen | Harmonic Mean |
|---|---|---|---|
| Manhattan distance | 0.74 | 0.48 | 0.582 |
| Chebyshev distance | 0.61 | 0.57 | 0.589 |
| Lance and Williams distance | 0.54 | 0.54 | 0.540 |
| Euclidean distance | **0.80** | 0.65 | 0.717 |
| Weighted Euclidean distance | **0.80** | **0.68** | **0.735** |

truth labels exceeds three. This shows that the proposed method is effective in challenging setups where multiple seen and unseen classes are predicted simultaneously.

We carried out feature visualization for various diseases and contrast it with the pathological location bounding boxes for those diseases that were provided in [54]. As shown in

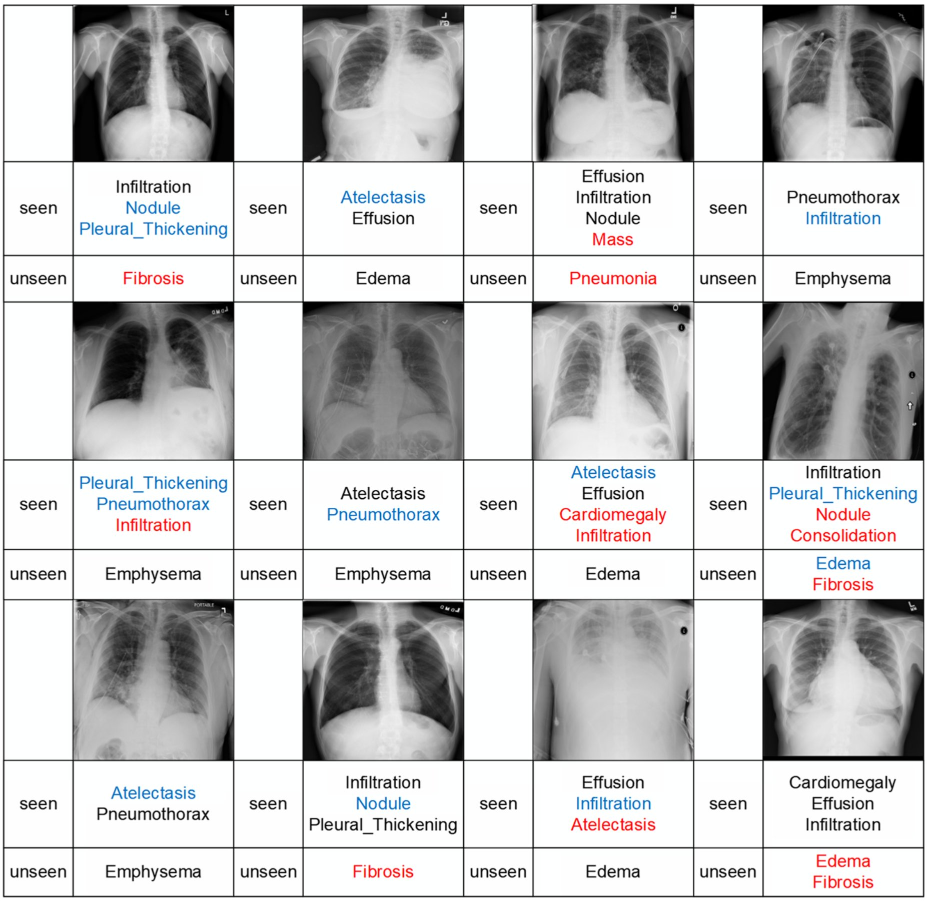


**Fig. 6** Results of the first three predictions for each image in the test set. Black represents true positives, red represents false negatives, and blue represents false positives. Black and red labels are combined to represent real labels

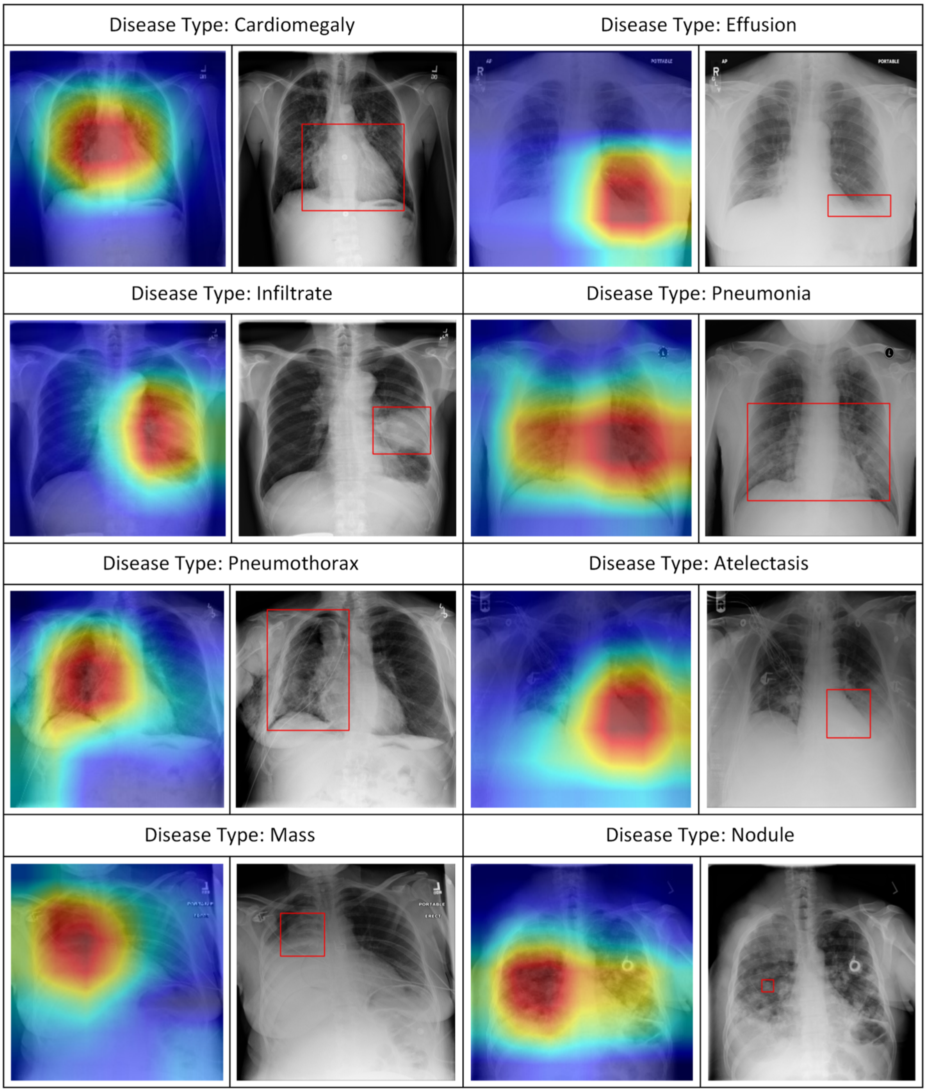


**Fig. 7** Visualization of the features of different diseases

Fig. 7, the darker the color is, the larger the value is. It can be considered that when the network predicts different diseases, the red highlighted area may be regarded as its primary criterion. It should be noted that since the image is cropped during feature extraction, the heatmap may have some deviations from the original image annotation box.

### 4.2.2 Ablation experiment

We performed ablation experiments. In all ablation experiments, we set the initial learning rate is set to 0.0001, the weight decay is 0.0003, and the batch sizes are 32. First, as shown in Table 4, the effect of $\alpha$ alone on the framework's overall performance is examined. Then

**Table 4** The effect of parameter $\alpha$ on the model

| $\alpha$ | $\beta$ | Seen | Unseen | Harmonic mean |
|---|---|---|---|---|
| 0 | 0 | 0.79 | 0.63 | 0.701 |
| 1 | 0 | **0.80** | **0.65** | **0.717** |
| 2 | 0 | 0.79 | **0.65** | 0.713 |
| 3 | 0 | 0.79 | 0.61 | 0.688 |

**Table 5** The effect of parameter $\beta$ on the model

| $\alpha$ | $\beta$ | Seen | Unseen | Harmonic mean |
|---|---|---|---|---|
| 0 | 0.00 | 0.79 | 0.63 | 0.701 |
| 0 | 0.01 | 0.79 | **0.69** | **0.736** |
| 0 | 0.02 | 0.79 | 0.66 | 0.719 |
| 0 | 0.03 | 0.79 | 0.67 | 0.725 |
| 1 | 0.01 | **0.80** | 0.62 | 0.698 |

**Table 6** Results of ablation experiments

| $\alpha$ | $\beta$ | Seen | Unseen | Harmonic mean |
|---|---|---|---|---|
| | | 0.79 | 0.63 | 0.701 |
| ✓ | | 0.79 | 0.65 | 0.713 |
| | ✓ | 0.79 | 0.67 | 0.725 |
| ✓ | ✓ | **0.80** | **0.68** | **0.735** |

Blank space represents 0

**Table 7** Experimental results of different backbone networks

| Feature extractor | Seen | Unseen | Harmonic mean |
|---|---|---|---|
| AlexNet [32] | 0.73 | 0.63 | 0.676 |
| VGG16 [52] | 0.74 | 0.61 | 0.668 |
| GoogleNet [53] | 0.78 | 0.65 | 0.709 |
| ResNet18 [20] | 0.78 | 0.67 | 0.720 |
| DenseNet121(ours) | **0.80** | **0.68** | **0.735** |

the impact of $\beta$ alone on the overall performance of the framework is analyzed, as shown in Table 5. The combined effect of $\alpha$ and $\beta$ on the framework's overall performance is shown in Table 6.

It can be seen from Tables 4 and 5 that the parameter $\alpha$ has a greater effect on the seen class, while the parameter $\beta$ can promote the classification accuracy of the unseen class. Table 5 displays the outcomes of our rerun experiment using the top-performing parameters $\alpha = 1$ and $\beta = 0.01$. Since the seen classes are known by the model during training, the model over-learning the seen class features will affect the classification performance of the unseen classes. Combined with the results in Figs. 4 and 5, we considered $\alpha = 2$ and $\beta = 0.03$ to be the best parameters for the experiment, and we further conducted ablation experiments, and the results are shown in Table 6.

In Table 6, '✓' represents $\alpha = 2$ or $\beta = 0.03$. According to the results of the ablation experiment, both parameter $\alpha$ and $\beta$ play a role in promoting the model, especially for unseen classes, and their contribution is greater when used together than when used alone.

As shown in Table 7, we also conduct experiments with different feature extractors (backbone networks) to demonstrate the efficacy of our model. The model's parameters remain unchanged, except for the feature extractor. DenseNet can comprehensively employ shallow features with low complexity, and it is easier to obtain a smooth decision function with superior generalization performance, so it can obtain more discriminative features than other models.

## 5 Discussion

Recently, numerous methods for multi-label zero-shot image recognition have been proposed by researchers. Among them, most of the current works use multimodal data for auxiliary research, but there are many difficulties when dealing with zero-shot images, especially the existence of multiple labels. In order to overcome the difficulties and improve the accuracy of multi-label zero-shot image recognition, this method uses fewer constraints and proposes a weighted focused metric loss function, which achieves greater accuracy. Evidently, metric learning has great potential in ZSL and can solve practical application problems.

Our future work will consist of refining our algorithm and employing novel deep learning techniques for multi-label few-shot rare disease diagnosis in order to improve disease detection capabilities.

## 6 Conclusion

In this paper, we design a method based on cross-modal deep metric learning to solve the multi-label zero-shot chest X-ray image classification problem and name the model CM-DML-GZSL. The three components of this model are a visual feature extractor, a fixed semantic feature extractor, and a deep regression module. In addition, we designed a weighted focused Euclidean distance metric loss function to increase the weight of hard samples and help the model to classify. Experiments validate our model and 3 sets of comparative experiments by using the large public dataset ChestX-ray14 dataset. According to the experimental results, our model achieves significant improvements on many individual classes. This study shows that deep metric learning has great potential for solving zero-shot problems, using simple and low-complexity metric losses to achieve better results than complex models. In the future, we will further study the detection problem of multi-label zero-shot chest X-ray images by combining convolution and Transformer. The method used will still be metric learning.

**Acknowledgements** In this paper, the research was sponsored by the National Natural Science Foundation of China (61272315), the Natural Science Foundation of Zhejiang Province (LY21F020028), and the Funded Project of Zhejiang Province University Student Scientific Research and Innovation Activities Plan (No.2021R409054).

**Funding** The National Natural Science Foundation of China (61272315), the Natural Science Foundation of Zhejiang Province (LY21F020028), and the Funded Project of Zhejiang Province University Student Scientific Research and Innovation Activities Plan (No.2021R409054).

**Data Availability** The datasets, models, and/or code generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

## Declarations

**Conflict of Interests** The authors declare that they have no conflict of interests.

## References

1. Akata Z, Perronnin F, Harchaoui Z, Schmid C (2013) Label-embedding for attribute-based classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 819–826
2. Akata Z, Reed S, Walter D, Lee H, Schiele B (2015) Evaluation of output embeddings for fine-grained image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2927–2936
3. Boutell MR, Luo J, Shen X, Brown CM (2004) Learning multi-label scene classification. Pattern Recogn 37(9):1757–1771
4. Cao X, Ge Y, Li R, Zhao J, Jiao L (2019) Hyperspectral imagery classification with deep metric learning. Neurocomputing 356:217–227
5. Chen ZM, Wei XS, Wang P, Guo Y (2019) Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5177–5186
6. Chen T, Xu M, Hui X, Wu H, Lin L (2019) Learning semantic-specific graph representation for multi-label image recognition. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 522–531
7. Cheng G, Yang C, Yao X, Guo L, Han J (2018) When deep learning meets metric learning: remote sensing image scene classification via learning discriminative CNNs. IEEE Trans Geosci Rem Sens 56(5):2811–2821
8. Cheng G, Li R, Lang C, Han J (2021) Task-wise attention guided part complementary learning for few-shot image classification. Sci Chin Inform Sci 64(2):1–14
9. Feng S, Fu P, Zheng W (2018) A hierarchical multi-label classification method based on neural networks for gene function prediction. Biotechnol Biotechnol Equip 32(6):1613–1621
10. Fu Y, Yang Y, Hospedales T, Xiang T, Gong S (2015) Transductive multi-label zero-shot learning. arXiv:1503.07790
11. Fürnkranz J, Hüllermeier E, Loza Mencía E, Brinker K (2008) Multilabel classification via calibrated label ranking. Mach Learn 73(2):133–153
12. Gabruseva T, Poplavskiy D, Kalinin A (2020) Deep learning for automatic pneumonia detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 350–351
13. Gaure A, Gupta A, Verma VK, Rai P (2017) A probabilistic framework for zero-shot multi-label learning. In: The Conference on Uncertainty in Artificial Intelligence (UAI), vol 1, p 3
14. Gouk H, Pfahringer B, Cree M (PMLR) Learning distance metrics for multi-label classification. In: Asian Conference on machine learning, pp 318–333
15. Gu K, Xia Z, Qiao J, Lin W (2019) Deep dual-channel neural network for image-based smoke detection. IEEE Trans Multimed 22(2):311–323
16. Gu K, Zhang Y, Qiao J (2020) Ensemble meta-learning for few-shot soot density recognition. IEEE Trans Industr Inform 17(3):2261–2270

17. Gu K, Liu H, Xia Z, Qiao J, Lin W, Thalmann D (2021) PM2.5 Monitoring: use information abundance measurement and wide and deep learning. IEEE Trans Neural Netw Learn Syst 32(10):4278–4290
18. Gupta A, Narayan S, Khan S, Khan FS, Shao L, van de Weijer J (2021) Generative multi-label zero-shot learning. arXiv:2101.11606
19. Hayat N, Lashen H, Shamout FE (2021) Multi-label generalized zero shot learning for the classiffcation of disease in chest radiographs. In: Machine learning for healthcare conference. PMLR, pp 461–477
20. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
21. Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504–507
22. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4700–4708
23. Huang H, Chen Y, Tang W, Zheng W, Chen QG, Hu Y, Yu P (2020) Multi-label zero-shot classification by learning to transfer from external knowledge. arXiv:2007.15610
24. Huang S, Lin J, Huangfu L (2020) Class-prototype discriminative network for generalized zero-shot learning. IEEE Signal Process Lett 27:301–305
25. Huynh D, Elhamifar E (2020) A shared multi-attention framework for multi-label zero-shot learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8776–8786
26. Ji Z, Cui B, Li H, Jiang YG, Xiang T, Hospedales T, Fu Y (2020) Deep ranking for image zero-shot multi-label classification. IEEE Trans Image Process 29:6549–6560
27. Ji Z, Wang H, Pang Y, Shao L (2020) Dual triplet network for image zero-shot learning. Neurocomputing 373:90–97
28. Jiang H, Wang R, Shan S, Chen X (2019) Adaptive metric learning for zero-shot recognition. IEEE Signal Process Lett 26(9):1270–1274
29. Jiang W, Huang K, Geng J, Deng X (2020) Multi-scale metric learning for few-shot learning. IEEE Trans Circuits Syst Video Technol 31(3):1091–1102
30. Karagoz GN, Yazici A, Dokeroglu T, Cosar A (2021) A new framework of multi-objective evolutionary algorithms for feature selection and multi-label classification of video data. Int J Mach Learn Cybern 12(1):53–71
31. Kodirov E, Xiang T, Gong S (2017) Semantic autoencoder for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3174–3183
32. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25
33. Lampert CH, Nickisch H, Harmeling S (2009) Learning to detect unseen object classes by between-class attribute transfer. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, pp 951–958
34. Lampert CH, Nickisch H, Harmeling S (2013) Attribute-based classification for zero-shot visual object categorization. IEEE Trans Pattern Anal Mach Intell 36(3):453–465
35. Lang C, Cheng G, Tu B, Han J (2022) Learning what not to segment: a new perspective on few-shot segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8057–8067
36. Lee CW, Fang W, Yeh CK, Wang YCF (2018) Multi-label zero-shot learning with structured knowledge graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1576–1585
37. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234–1240
38. Li C, Liu C, Duan L, Gao P, Zheng K (2019) Reconstruction regularized deep metric learning for multi-label image classification. IEEE Trans Neural Netw Learn Syst 31(7):2294–2303
39. Lu J, Hu J, Tan YP (2017) Discriminative deep metric learning for face and kinship verification. IEEE Trans Image Process 26(9):4269–4282
40. Lu X, Wang W, Ma C, Shen J, Shao L, Porikli F (2019) See more, know more: unsupervised video object segmentation with co-attention siamese networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3623–3632
41. Lu X, Wang W, Shen J, Crandall D, Luo J (2020) Zero-shot video object segmentation with co-attention siamese networks. IEEE Transactions on Pattern Analysis and Machine Intelligence
42. Mahapatra D, Bozorgtabar B, Ge Z (2021) Medical image classification using generalized zero shot learning. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3344–3353
43. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv:1301.3781

44. National Institutes of Health (2017) NIH Clinical Center provides one of the largest publicly available chest x-ray datasets to scientific community
45. Peng J, Bu X, Sun M, Zhang Z, Tan T, Yan J (2020) Large-scale object detection in the wild from imbalanced multi-labels. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9709–9718
46. Romera-Paredes B, Torr P (PMLR) An embarrassingly simple approach to zero-shot learning. In: International conference on machine learning, pp 2152–2161
47. Salim I, Hamza AB (2021) Ridge regression neural network for pediatric bone age assessment. Multimed Tools Appl 80(20):30461–30478
48. Sariyildiz MB, Cinbis RG (2019) Gradient matching generative networks for zero-shot learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2168–2178
49. Sharma H, Jain JS, Bansal P, Gupta S (2020) Feature extraction and classification of chest x-ray images using cnn to detect pneumonia. In: 2020 10th International conference on cloud computing, data science and engineering (confluence). IEEE, pp 227–231
50. Shen Y, Qin J, Huang L, Liu L, Zhu F, Shao L (2020) Invertible zero-shot recognition flows. In: European conference on computer vision. Springer, Cham, pp 614–631
51. Shigeto Y, Suzuki I, Hara K, Shimbo M, Matsumoto Y (2015) Ridge regression, hubness, and zero-shot learning. In: Joint European conference on machine learning and knowledge discovery in databases. Springer, Cham, pp 135–151
52. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
53. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
54. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM (2017) Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2097–2106
55. Wang W, Lu X, Shen J, Crandall DJ, Shao L (2019) Zero-shot video object segmentation via attentive graph neural networks. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9236–9245
56. Xian Y, Akata Z, Sharma G, Nguyen Q, Hein M, Schiele B (2016) Latent embeddings for zero-shot classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 69–77
57. Xian Y, Lorenz T, Schiele B, Akata Z (2018) Feature generating networks for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5542–5551
58. Xie GS, Liu L, Jin X, Zhu F, Zhang Z, Qin J, Shao L (2019) Attentive region embedding network for zero-shot learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9384–9393
59. Xu W, Xian Y, Wang J, Schiele B, Akata Z (2020) Attribute prototype network for zero-shot learning. Adv Neural Inf Process Syst 33:21969–21980
60. Yan Z, Liu W, Wen S, Yang Y (2019) Multi-label image classification by feature attention network. IEEE Access 7:98005–98013
61. Yu B, Tao D (2019) Deep metric learning with tuplet margin loss. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6490–6499
62. Yu Y, Ji Z, Guo J, Zhang Z (2018) Zero-shot learning via latent space encoding. IEEE Trans Cybern 49(10):3755–3766
63. Zhang F, Shi G (2019) Co-representation network for generalized zero-shot learning. In: International conference on machine learning. PMLR, pp 7434–7443
64. Zhang ML, Zhou ZH (2013) A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng 26(8):1819–1837
65. Zhang L, Xiang T, Gong S (2017) Learning a deep embedding model for zero-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2021–2030
66. Zhang M, Li C, Wang X (2019) Multi-view metric learning for multi-label image classification. In: 2019 IEEE International Conference on Image Processing (ICIP). IEEE, pp 2134–2138
67. Zunair H, Hamza AB (2020) Melanoma detection using adversarial training and deep transfer learning. Phys Med Biol 65(13):135005
68. Zunair H, Hamza AB (2021) Synthesis of COVID-19 chest X-rays using unpaired image-to-image translation. Soc Netw Anal Mining 11(1):1–12

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Affiliations

**Yufei Jin[1,2] · Huijuan Lu[1,2] · Zhao Li[3] · Yanbin Wang[4]**

Yufei Jin
s20030812005@cjlu.edu.cn

Zhao Li
zhao_li@zju.edu.cn

Yanbin Wang
wangyanbin15@mails.ucas.ac.cn

[1] China JiLiang University, Hangzhou, 310018, Zhejiang, China

[2] Key Laboratory of Electromagnetic Wave Information Technology and Metrology of Zhejiang Province, Hangzhou, 310018, Zhejiang, China

[3] Zhejiang University, Hangzhou, 310018, Zhejiang, China

[4] College of Computer Science and Technology, Zhejiang University, Hangzhou, 310018, Zhejiang, China