Information Processing and Management 60 (2023) 103258

Contents lists available at ScienceDirect

Information Processing and Management

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

IGCNN-FC: Boosting interpretability and generalization of convolutional neural networks for few chest X-rays analysis

Mengmeng Zhan [a], Xiaoshuang Shi [a,*], Fangqi Liu [a], Rongyao Hu [b]

[a] School of Computer Science and Technology, University of Electronic Science and Technology of China, Chengdu 611731, China
[b] Massey University Albany Campus, Auckland 0745, New Zealand

ARTICLE INFO

Keywords:
CNNs
Data scarcity
Interpretability
Chest X-ray

ABSTRACT

Computer-aided diagnosis (CAD) with convolutional neural networks (CNNs) has been widely applied to assist doctors in medical image analysis. However, most of them often encounter two obstacles: (1) Data scarcity, because the advanced performance of CNNs heavily depends on a large amount of data, especially high-quality annotated ones. (2) Interpretability, CNNs cannot directly provide evidence related to the decision-making process to support their diagnosis results. To overcome these two obstacles, we propose an interpretable deep learning framework based on CNNs. Specifically, we introduce a multi-scale loss-based attention to leverage the mid- and high-level features to mine significant features for decision-making. Additionally, to better explore the semantic knowledge from training data, we utilize the mixup method to produce more annotated training images. Moreover, to boost model generalization capability, we employ the self-distillation to learn the knowledge generated from previous training epochs. Experiments on two benchmark Chest X-ray datasets demonstrate the effectiveness of the proposed framework with superior performance over recent SOTA methods, with boosting model interpretability.

1. Introduction

Pulmonary diseases, such as tuberculosis (Sterling, Njie, Zenner, et al., 2020), pleural effusion (Lu, Luo, Chen, Zhuansun, et al., 2020) and coronavirus disease (covid-19) (Zhu, Song, Shi, et al., 2021), are serious threat to human life and health. Early screening can effectively intervene the disease progression and boost the success rate of treatment. Chest X-ray is one of the most widely used medical imaging analysis techniques in clinical diagnosis. It is necessary for radiologists to analyze Chest X-ray data so as to accurately diagnose and treat many lung diseases (Silva, Poellinger, Cardoso, & Reyes, 2020). However, manually checking Chest X-rays is laborious, time-consuming and error-prone. Hence, many CAD systems have been developed to assist radiologists to reduce their workloads. Recently, because CNNs can automatically extract powerful feature representations from large-scale data, they have been widely applied to various CAD systems (Chen, Lu, Chen, Williamson, & Mahmood, 2021), including the automatic classification of pulmonary tuberculosis (Wu, Wang, & Wu, 2017), the prediction of covid-19 (Zhu et al., 2021), etc. The success of CNNs depends on large-scale high-quality annotated data (Thulasidasan, Chennupati, Bilmes, Bhattacharya, & Michalak, 2019), however, it is laborious, time-consuming and expensive to attain a large number of annotated medical images (Kaissis, Ziller, Passerat-Palmbach, et al., 2021). Additionally, CNNs cannot interpret the process of their decision-making. But it is essential and challenging for CNNs to utilize only a small amount of medial data to obtain interpretable diagnosis (Barnett et al., 2021).

* Corresponding author.
E-mail address: xsshi2021@uestc.edu.cn (X. Shi).

https://doi.org/10.1016/j.ipm.2022.103258
Received 12 May 2022; Received in revised form 13 December 2022; Accepted 19 December 2022
Available online 9 January 2023

CNNs often consist of a huge number of parameters, hence, they are easily overfitting on insufficient annotated training data, leading to poor generalization capability. To overcome this obstacle, one popular strategy is to automatically annotate a large amount of data by automation or a crowdsourcing platform. However, this strategy usually generates severe label noise, thereby restricting its applicability on medical images (Kaissis et al., 2021). Another popular strategy is transfer learning, which usually has two stages: pre-training by a large auxiliary dataset and fine-tuning by a small target dataset (Nguyen, Luu, Pham, Rakhimkul, & Yoo, 2021; Shang, Xie, et al., 2021). Nevertheless, this strategy requires the target and auxiliary datasets to have similar distributions. Additionally, pre-training often consumes massive storage and calculation costs (He, Girshick, & Dollár, 2019). The third popular strategy is data augmentation, which can produce more variants of existing scarce training data to enlarge the training set. This strategy can reduce the variance of the mapping learned by deep learning models, and has been effectively applied to various medical diagnosis systems (Noguchi, Nishio, Yakami, Nakagomi, & Togashi, 2020).

The second notorious obstacle for CNNs is lack of interpretability, which severely restricts their applications in clinical medical diagnosis. The interpretability of models has received a lot of attention (Shi, Xing, Xu, et al., 2020; Zhu, Ma, Yuan, & Zhu, 2022; Zhu, Zhang, Zhu, Zhu, & Gao, 2020). For example, Zhu et al. (2022) proposed a dynamic feature selection method to select the most differentiated brain regions for the diagnosis of Alzheimer's disease. To overcome this obstacle, attention mechanisms have been widely studied to boost model interpretation capability (Shi, Xing, Xu, et al., 2020), because they can better interpret the decision-making process and more accurately discover the significant features than visual interpretation methods, like Gradient-weighted Class Activation Mapping (Grad-CAM) (Zhang et al., 2020). However, attention mechanisms often lead to a low recall of significant features, thereby possibly decreasing the model classification performance. To address this limitation, recently, loss-based attention mechanism (Shi, Xing, Xie, et al., 2020; Shi, Xing, Xu, et al., 2020) is proposed to maximally select salient regions and meanwhile classify images. But it only considers high-level features in CNNs and neglects the mid-level feature representations, since many literatures indicate that mixing features of different levels can significantly boost the model performance (Cao, Puy, Boulch, & Marlet, 2021; Deng, Wang, Liu, Liu, & Jiang, 2021; Qiao, Chen, & Yuille, 2021).

Based on the aforementioned observations, in this paper, we propose a novel ***i***nterpretable and ***g***eneralization ***c***onvolutional ***n***eural ***n***etworks for ***f***ew ***c***hest X-rays analysis, namely **IGCNN-FC**. Specifically, we embed two layers of loss-based attention into CNN with simultaneously considering the mid- and high-level features, so as to better mine the significant features for decision-making. Additionally, we employ the popular data augmentation method mixup to produce more training data to alleviate data scarcity. To further enhance the model generalization capability, we introduce a self-distillation method to employ the knowledge learned by model training. In summary, our major contributions are listed as follows:

- We propose a novel interpretable CNN for few Chest X-rays analysis, by efficiently integrating loss-based attention, mixup and self-distillation into a unified deep framework, which can mine significant features to interpret diagnosis results using only few annotated data.
- We design a multi-scale loss-based attention with simultaneously employing mid- and high-level features in the network to boost model performance and interpretability.
- Extensive experiments on two popular Chest X-ray datasets demonstrate that the proposed IGCNN-FC method has superior performance over recent SOTA methods with better interpretability.

2. Relation work

2.1. Few-shot learning

2.1.1. Meta-learning based methods

Meta-learning methods mainly utilize the idea of learning how to learn to make the model acquire a better learning ability (Huisman, Van Rijn, & Plaat, 2021). Specifically, meta-learning learns meta-knowledge from many tasks and uses the obtained meta knowledge to guide the model to learn faster in new tasks that contain only a few data (Shu, Cao, Wang, Wang, & Long, 2021). For example, MAML (Finn, Abbeel, & Levine, 2017) proposed learning how to obtain a good initialization. Meta-SGD (Li, Zhou, Chen, & Li, 2017) further learns model optimization's direction and learning rate based on MAML. Meta-learning methods usually employ complex model structures or second-order gradient optimization, so it is difficult to train and learn in the process of actual optimization learning (Sun et al., 2021). Therefore, the problem of meta-learning efficiency in small sample learning needs to be solved.

2.1.2. Metric-learning based methods

Metric-based methods aim to determine its category by measuring the similarity between the query and support samples (Singh, Hie, Narayan, & Berger, 2021). The framework for this type of method usually has two modules: embedded module and measurement module (Kim, Kim, Cho, & Kwak, 2021). First, samples are embedded into vector space through the embedded module. Then, similarity scores are given according to the measurement module. Koch, Zemel, Salakhutdinov, et al. (2015) first proposed to use the Siamese-net for few-shot image recognition. Vinyals, Blundell, Lillicrap, et al. (2016) introduced a matching network using the attention mechanism based on LSTM to classify samples by comparing the distance between different category representations. In order to further solve the problem of few samples, Snell, Swersky, and Zemel (2017) proposed a prototypical network to measure the distance between different samples and prototype centers. The above model calculates similarity based on a distance function but does not consider the non-linear distance problem. Based on the prototype network, Sung, Yang, Zhang, et al. (2018) proposed

a Relation Network employ CNN to better measure the distance between samples, making the measurement method more reliable and flexible. Compared with the meta-learning method, measurement learning is faster, easier to optimize, and gradually becomes an essential branch of few-shot learning (Cen, Yun, Cai, Wang, & Liu, 2021). However, it lacks sufficient persuasion in explaining the model results.

2.1.3. Data augmentation

The data augmentation method attempts to increase the training samples to improve the performance of few-shot learning. For image data, data augmentation can be realized through basic translation, rotation, flipping, and other operations (Osahor & Nasrabadi, 2022). Because the types of data generated by basic image transformation may be limited, in recent years, more and more methods have focused on mixing existing annotation data to generate new data (Chen et al., 2019; Hariharan & Girshick, 2017; Thulasidasan et al., 2019). This method is easier to implement than the data augmentation method based on basic image transformation, and the generated samples have greater variability. SGM (Hariharan & Girshick, 2017) proposed a new strategy to generate hallucinate additional samples to alleviate the problem of data scarcity. Peng et al. (2022) and Yuan, Zhong, Lei, Zhu, and Hu (2021) introduced a reverse graph model in the original data space to obtain high-quality node data. Additionally, Chen et al. (2019) proposed IDeMeNet, which obtains many training samples through adaptive fusion support set samples. Mixup (Thulasidasan et al., 2019) generated a large number of enhanced images by mixing the two images using linear interpolation. Cutmix (Yun et al., 2019) further developed and proposed to cut and paste a part of a randomly selected image into another image. Further, improve the variability of synthetic samples. Supermix (Dabouei, Soleymani, Taherkhani, & Nasrabadi, 2021) mixed salient areas of different images to generate new images. Stylemix (Hong, Choi, & Kim, 2021) utilized image style and contented to improve the process of image interpolation. In medical image analysis tasks, Cyclemix (Zhang & Zhuang, 2022) adopted a hybrid strategy and designed a random occlusion method to generate many medical images. Because data augmentation is simple and applicable, it has become a popular method in few-shot learning.

2.2. Knowledge distillation

Knowledge distillation aims to improve the performance of student models by using the knowledge acquired from a well-generalized large model (teacher) to guide the training of small models. Recently, various extension methods of knowledge distillation have been proposed (Beyer et al., 2022; Gou, Yu, Maybank, & Tao, 2021; Wang & Yoon, 2021). For example, Kang, Zhang, Zhang, Sun, and Zheng (2021) proposed an instance knowledge distillation framework and extended it to target detection. Shang, Duan, Zong, Nie, and Yan (2021) suggested that Lipschitz continuity can be used to guide the knowledge distillation framework and improve student model performance. Despite complex teacher models can improve the generalization performance of student models, pre-training requires high time and economic cost. Therefore, self-distillation using the same network for teacher and student models utilizes its own knowledge to learn without high training costs . Andonian, Chen, and Hamid (2022) proposed to solve the limitations of info-NCE loss training on noise cross-modal data by progressive self-distillation. Yoon, Kang, and Cho (2022) applied self-distillation to reduce the differences between the two domains in domain adaptation.

2.3. Attention mechanism

In visual cognition, human beings selectively devote themself to some information and ignore other visible information. Inspired by this, attention mechanism is widely used in computer vision (Guo et al., 2022; Wang, Zhang, Kan, Shan, & Chen, 2020), graph learning (Gan et al., 2022; Mo, Peng, Xu, Shi, & Zhu, 2022; Song et al., 2022; Xue et al., 2022) and other tasks (Cai et al., 2020; Ma et al., 2022). For example, CAN (Hou, Chang, Ma, Shan, & Chen, 2019) constructed the relationship between supporting and querying images through a cross-attention mechanism. Xu et al. (2022) focused on vital information in different modes to improve the performance of multimodal clustering. In addition to boosting the model performance, the attention mechanism also makes the model interpretable (Shi, Xing, Xie, et al., 2020; Song et al., 2022). For example, Song et al. (2022) proposed to capture the important relationship between cross-concept exercises in knowledge tracking through the attention mechanism and improved the model interpretability. Recently, loss-based attention mechanism (Shi, Xing, Xie, et al., 2020; Shi, Xing, Xu, et al., 2020) was proposed to maximally select salient regions and meanwhile classify images. Inspired by this, we tailor a novel multi-scale attention mechanism for solving the medical image analysis.

3. Method

We present the overflow of the proposed framework in Fig. 1, and introduce its four major modules: image classification, multi-scale loss-based attention, mixup and self-distillation in the following.

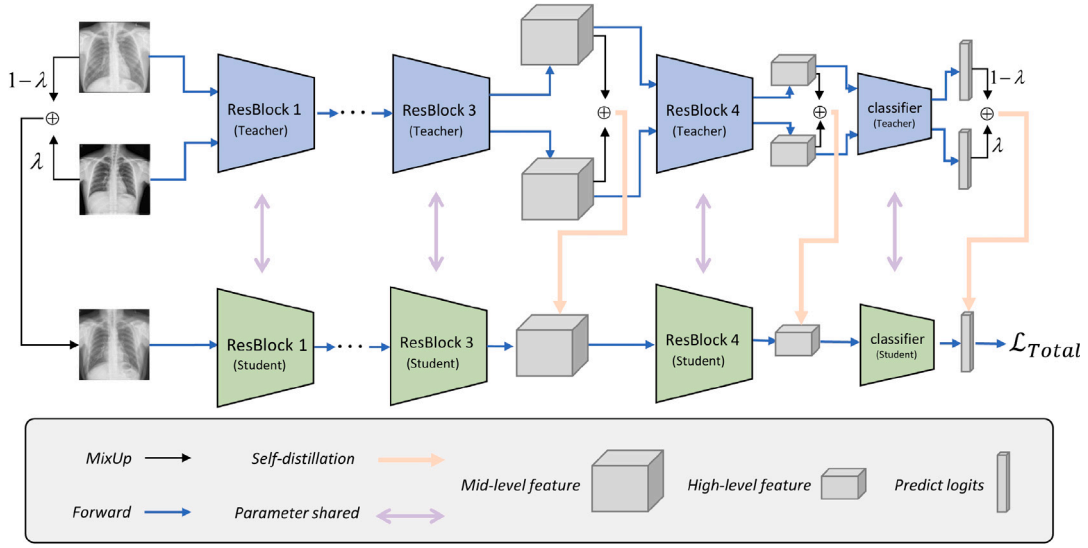


Fig. 1. The framework of the proposed IGCNN-FC. We utilize ResNet18 as the backbone network.

3.1. Image classification

Given training data $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^{C\times W\times H}$ is the ith image and N is the total number of images, C, W and H represent the number of channels, the image width and height, respectively. $\mathbf{y}_i \in \{0,1\}^K$ is a one-hot vector generated by the label of $\mathbf{x}_i$, where K is the number of classes. Let $f(\cdot)$ represent the backbone network, $\mathbf{z}_i = f(\mathbf{x}_i) \in \mathbb{R}^K$ be the final output of the network, $\mathbf{p}_i = s(\mathbf{z}_i)$ denote the estimated class probability, where $s(\cdot)$ is a softmax function. Then, we can utilize the cross-entropy loss to train the model for image classification, i.e.,

$$\mathcal{L}_1 = -\sum_{i=1}^{N} \mathbf{y}_i log(\mathbf{p}_i). \tag{1}$$

By minimizing Eq. (1), we can obtain the optimal parameters Θ of the backbone network.

3.2. Multi-scale loss-based attention

To boost the interpretability of CNNs by mining the significant features for decision-making, recently, Shi, Xing, Xu, et al. (2020) proposed a loss-based attention mechanism to simultaneously learn the prediction of features and their weights. Suppose that one medical image $\mathbf{x}_i$ is divided into M patches, upon which the network can obtain M_1 high-level features. Let $\mathbf{h}_{im}$ represent the mth feature in the representation, $\mathbf{u}_{im} = \mathbf{h}_{im}\theta \in \mathbb{R}^K$ be its logits, and θ represent the parameters of the fully connected layer. Then, the loss-based attention mechanism is formulated as:

$$\begin{aligned} \alpha_{im} &= \frac{\mathbf{u}_{im}[y_i]}{\sum_{m=1}^{M_1} \mathbf{u}_{im}[y_i]}, \\ \alpha_{im} &= \frac{\max(\alpha_{im} - \frac{\xi}{M_1}, 0)}{\sum_{m=1}^{M_1} \max(\alpha_{im} - \frac{\xi}{M_1}, 0)}, \\ \mathbf{h}_{im} &\leftarrow \alpha_{im}\mathbf{h}_{im}, \\ \mathbf{z}_i &= \sum_{m=1}^{M_1} \mathbf{h}_{im}\theta, \end{aligned} \tag{2}$$

where α_{im} is the attention weight of the mth feature in the ith image $\mathbf{x}_i$, and ξ is a nonnegative parameter to remove the trivial features. However, Eq. (2) only takes into account high-level feature representations and neglects the low-level ones. Much of the literature has demonstrated that combining the features of different levels can obtain richer semantic information than only using high-level features, thereby boosting model performance (Deng et al., 2021; Qiao et al., 2021). Therefore, we design a multi-scale loss-based attention to boost model interpretability and classification performance by combining different levels of features. Specifically, we employ the features extracted by the third and fourth Resblock in ResNet18 as the mid- and high-level ones, respectively. Because mid- and high-level features have different scales, the loss-based attention simultaneously handling mid- and high-level features is named as multi-scale loss-based attention. By using Eq. (2), it is easy to obtain the weights of mid- and high-level features. Then, we design the multi-scale loss function as follows:

$$\mathcal{L}_2 = -\sum_{i=1}^{N} \left(\sum_{m=1}^{M_1} \alpha_{im}\mathbf{y}_i log\left(\mathbf{p}_{im}\right) + \sum_{m=1}^{M_2} \beta_{im}\mathbf{y}_i log\left(\mathbf{q}_{im}\right) \right), \tag{3}$$

where $\mathbf{p}_{im} = s(\mathbf{u}_{im})$, $\mathbf{u}_{im}$ is the logits generated by high-level features; $\mathbf{q}_{im} = s(\mathbf{v}_{im})$, $\mathbf{v}_{im} \in \mathbb{R}^K$ is the logits generated by mid-level features, M_2 is the number of mid-level features extracted by the third Resblock, similar to α_{im}, β_{im} is the weight of the mth mid-level feature.

3.3. Mixup

To alleviate the problem of data scarcity, we adopt mixup (Thulasidasan et al., 2019) to augment training data and their targets, because it can boost the diversity of data by linearly mixing two images and their corresponding true class probabilities. Specifically, suppose that $\mathbf{x}_i$ and $\mathbf{x}_j$ denote the ith and j^{th} image, respectively, $\mathbf{p}_i$ and $\mathbf{p}_j$ are true class probabilities of $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. $\mathbf{y}_i$ and $\mathbf{y}_j$ are two one-hot label vectors of $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. Mixup employs linear interpolation to expand the training distribution as follows:

$$\begin{aligned} \lambda &\sim Beta(\mu, \mu), \\ \mathbf{x} &= \lambda \mathbf{x}_i + (1-\lambda)\mathbf{x}_j, \\ \mathbf{y} &= \lambda \mathbf{y}_i + (1-\lambda)\mathbf{y}_j, \end{aligned} \tag{4}$$

where the merging coefficient λ is sampled from a Beta distribution parameterized by μ, $\mathbf{x}$ represents the mixed data and $\mathbf{y}$ denotes its one-hot label vector. Because Eq. (4) only mixes image-level knowledge, we extend it to leverage the patch-level knowledge as follows:

$$\begin{aligned} \mathbf{p}_m &= \lambda \mathbf{p}_{im} + (1-\lambda)\mathbf{p}_{jm}, \\ \mathbf{q}_m &= \lambda \mathbf{q}_{im} + (1-\lambda)\mathbf{q}_{jm}, \end{aligned} \tag{5}$$

where $\mathbf{p}_m$ and $\mathbf{q}_m$ denote the mixed class probability of high-level and mid-level features, respectively. $\mathbf{p}_{im}$ and $\mathbf{p}_{jm}$ represent the prediction class probability of the mth high-level feature in $\mathbf{x}_i$ and $\mathbf{x}_j$, respectively. Similarly, $\mathbf{q}_{im}$ and $\mathbf{q}_{jm}$ are obtained from the mth mid-level feature.

3.4. Self-distillation

To boost the generalization capability of CNNs, knowledge distillation (KD) is an effective and popular strategy to distill knowledge from a complex teacher model to a small student model through soft targets. Self-distillation is one method of KD to employ the identical teacher and student model for knowledge distillation. Recently, RFS (Tian, Wang, Krishnan, Tenenbaum, & Isola, 2020) adopts a two-stage self-distillation mechanism consisting of supervised training and self-distillation. Because supervised training and self-distillation are two individual stages during training, they consume more training cost. To this end, we propose a one-stage self-distillation mechanism that integrates supervised training and KD into a unified framework. Specifically, given a set $\mathcal{B}$ containing one mini-batch, where images are sampled from the training data, we can generate two sets $\mathcal{B}_1$ and $\mathcal{B}_2$ by augmenting each image in $\mathcal{B}$ through standard data augmentation. Additionally, by feeding the images in $\mathcal{B}_1$ and $\mathcal{B}_2$ into the teacher model, it is easy to obtain the prediction class probability of each image, its high-level and mid-level features, i.e., $\mathbf{p}_i$, $\mathbf{p}_{im}$, $\mathbf{q}_{im}$ for the image $\mathbf{x}_i \in \mathcal{B}_1$, and $\mathbf{p}_j$, $\mathbf{p}_{jm}$, $\mathbf{q}_{jm}$ for the image $\mathbf{x}_j \in \mathcal{B}_2$. Next, according to the mixup strategy, we first linearly mix the images in the two sets $\mathcal{B}_1$ and $\mathcal{B}_2$, and then feed the mixed images into the student model to obtain the prediction class probability of the image $\mathbf{x}$, and its high-level and mid-level features, i.e., $\mathbf{p}$, $\mathbf{p}_m$ and $\mathbf{q}_m$. After that, similar to Eq. (5), we mix the prediction class probability obtained by the teacher model as the target for the student model, i.e.,

$$\begin{aligned} \mathbf{p}^{t-1} &= \lambda \mathbf{p}_i + (1-\lambda)\mathbf{p}_j, \\ \mathbf{p}_m^{t-1} &= \lambda \mathbf{p}_{im} + (1-\lambda)\mathbf{p}_{jm}, \\ \mathbf{q}_m^{t-1} &= \lambda \mathbf{q}_{im} + (1-\lambda)\mathbf{q}_{jm}, \end{aligned} \tag{6}$$

where t is the number of training epochs, $\mathbf{p}^{t-1}$, $\mathbf{p}_m^{t-1}$ and $\mathbf{q}_m^{t-1}$ are the targets of $\mathbf{p}$, $\mathbf{p}_m$ and $\mathbf{q}_m$, respectively. Then, we employ the Kullback–Leibler (KL) divergence to distill knowledge from the teacher model as follows:

$$\mathcal{L}_3 = D_{KL}(\mathbf{p} \| \mathbf{p}^{t-1}) + D_{KL}(\mathbf{p}_m \| \mathbf{p}_m^{t-1}) + D_{KL}(\mathbf{q}_m \| \mathbf{q}_m^{t-1}), \tag{7}$$

where D_{KL} denotes the KL divergence.

Finally, to simultaneously take into account image classification, multi-scale loss-based attention, mixup, and self-distillation, the final total loss is:

$$\mathcal{L} = \mathcal{L}_1 + \omega(t)(\delta \mathcal{L}_2 + \gamma \mathcal{L}_3), \tag{8}$$

where δ and γ are constants to weight $\mathcal{L}_2$ and $\mathcal{L}_3$, respectively, $\omega(t)$ is a linear function, whose value is linearly changing with the number of training epochs t. For clarity, we present the detailed procedure of the proposed IGCNN-FC in Algorithm 1.

Algorithm 1 IGCNN-FC

Input: Training images $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^{N}$ and their one-hot labels $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^{N}$, the number of training epochs T, parameter α for MixUp, weight parameters δ, γ,
linear function $\omega(t)$,
stochastic input augmentation function: $h(\cdot)$,
stochastic teacher and student model with parameters Θ_t,Θ_s: $f(\cdot)$
Output: Student model parameters Θ_s

for t **in** $[1, T]$ **do**
  for each mini-batch $\mathcal{B}$ **do**
    $\mathcal{B}_1,\ \mathcal{B}_2 \leftarrow h\left(\mathbf{x}_i \in \mathcal{B}\right)$ ▷ *Data augmentation*
    $\mathbf{p}_i, \mathbf{p}_{im}, \mathbf{q}_{im} \leftarrow f\left(\Theta_t, \mathbf{x}_{i\in\mathcal{B}_1}\right)$
    $\mathbf{p}_j, \mathbf{p}_{jm}, \mathbf{q}_{jm} \leftarrow f\left(\Theta_t, \mathbf{x}_{j\in\mathcal{B}_2}\right)$
    $\mathbf{p}^{t-1}, \mathbf{p}_m^{t-1}, \mathbf{q}_m^{t-1} \leftarrow Eq.(6)$ ▷ *Mixup*
    $\mathbf{p}, \mathbf{p}_m, \mathbf{q}_m \leftarrow f\left(\Theta_s, \mathbf{x}_{i\in\mathcal{B}}\right)$
    $Loss \leftarrow$ Eq. (8) ▷ *Loss function*
    updating Θ_t,Θ_s using Adam;
  end for
  updating function $\omega(t)$; ▷ *Linearly ramp up $\omega(t)$ to its maximum value*
end for

4. Experiments

4.1. Datasets

Shenzhen set: Shenzhen dataset (Candemir, Jaeger, Palaniappan, et al., 2013; Jaeger, Karargyris, Candemir, et al., 2013) has 662 images, consisting of 326 normal and the rest are abnormal Chest X-rays (CXRs) with the manifestation of tuberculosis. For the Shenzhen dataset, we divide it into two cases: (1) 30% and 70% of the total number of images are used for training and testing, respectively; (2) 50% of the whole data are used for training and the remaining images are used for testing. **NIH set:** NIH Chest X-ray (Wang, Peng, Lu, et al., 2017) contains 112,120 frontal-view CXR images with fourteen diseases. Same as Bozorgtabar, Mahapatra, Vray, and Thiran (2020), we combine all labels of diseases into one category and then totally select 10,280 images, including 5170 normal and 5110 abnormal CXRs. Similarly, for the NIH dataset, we also divide it into two cases: 5% and 10% of the whole data are used for training, respectively, and the remaining ones are employed for testing. We repeat this process 5 times and report the average results.

We compare the proposed framework with six popular few shot learning methods and list their details as follows:

- ProtoNet (Snell et al., 2017) predicted the category of the test sample by calculating the Euclidean distance between the test sample and the center of each prototype.
- RelationNet (Hu, Gu, Zhang, Dai, & Wei, 2018) predicted the category of samples by using learnable modules instead of Euclidean distance.
- RFS (Tian et al., 2020) optimized the training process of the embedded model and used self-supervised knowledge distillation to improve the embedded model's performance further.
- Baseline (Chen, Liu, Kira, et al., 2018) fixed the pretrained feature extractor and finetuned the FC layer, and Baseline++ (Chen et al., 2018) replaced the top linear layer with a cosine classifier.
- ANIL (Raghu, Raghu, Bengio, & Vinyals, 2019) can be regarded as an improvement of MAML, and ANIL does not update all network layer parameters, but only the last layer network parameters.
- RENet (Kang, Kwon, Min, & Cho, 2021) jointed the self-correlation representation information of image data and cross-correlation attention mechanism to learn relational embeddings.

4.2. Implementation details

We trained our IGCNN-FC method using the PyTorch framework on an NVIDIA GTX 3090TI GPU (24 GB memory) and used Adam optimizer as the model optimization. We set the batch size and training epoch for all datasets as 32 and 100, respectively. Notably, we adapted the resolution of all images to 224 × 224 pixels with central cropping to improve efficiency. The hyper-parameter ξ is empirically set as 0.1 for the Shenzhen dataset and 0.05 for the NIH dataset. Meanwhile, the other hyper-parameters δ, γ are empirically set as [0.01, 0.05, 0.1, 0.5, 1] and [0.1, 1, 10, 50, 100], respectively. We performed a sensitivity analysis about hyper-parameters δ and γ in Fig. 3. Additionally, we adopt ResNet18 as the backbone network and the same data augmentation for the comparative methods. We evaluated all methods by using four popular metrics: classification accuracy (ACC), specificity (SPE), sensitivity (SEN) and the AUC score.

Table 1
Classification performance on the Shenzhen and NIH datasets with different rates of the whole data for training, respectively. We bold the best result in each group.

Model	Shenzhen dataset (30%)				NIH dataset (5%)			
	ACC	SPE	SEN	AUC	ACC	SPE	SEN	AUC
ProtoNet	82.67 ± 2.4	90.23 ± 2.1	74.75 ± 5.5	82.75 ± 1.6	81.17 ± 0.8	85.86 ± 2.4	76.48 ± 1.1	81.17 ± 0.7
RelationNet	80.63 ± 2.6	86.27 ± 4.8	74.55 ± 6.3	81.33 ± 2.2	81.41 ± 0.5	85.10 ± 2.8	**77.62 ± 2.4**	81.36 ± 0.5
Baseline	81.42 ± 2.4	89.93 ± 2.8	73.09 ± 4.9	81.51 ± 2.4	81.85 ± 0.7	86.44 ± 1.8	77.21 ± 0.7	81.82 ± 0.8
Baseline++	82.58 ± 1.9	88.51 ± 1.9	76.87 ± 5.0	82.69 ± 1.7	82.23 ± 0.7	88.04 ± 1.5	76.63 ± 0.4	82.34 ± 0.6
RFS	82.41 ± 0.6	89.38 ± 1.2	75.17 ± 2.3	82.27 ± 0.6	81.87 ± 0.5	87.50 ± 1.5	76.30 ± 1.4	81.84 ± 0.6
ANIL	82.41 ± 1.8	88.08 ± 2.5	**76.93 ± 3.8**	82.50 ± 1.7	82.12 ± 0.5	87.38 ± 1.7	77.13 ± 1.3	82.25 ± 0.4
RENet	82.04 ± 2.0	89.48 ± 1.5	74.30 ± 4.2	81.89 ± 2.0	82.11 ± 0.5	88.17 ± 1.5	76.12 ± 1.9	82.15 ± 0.4
IGCNN-FC	**83.00 ± 1.4**	**91.57 ± 1.0**	74.37 ± 2.8	**82.94 ± 1.4**	**83.10 ± 0.4**	**89.33 ± 2.0**	76.85 ± 2.4	**83.24 ± 0.4**
Model	Shenzhen dataset (50%)				NIH dataset (10%)			
	ACC	SPE	SEN	AUC	ACC	SPE	SEN	AUC
ProtoNet	83.75 ± 1.6	90.87 ± 3.2	76.00 ± 2.2	83.43 ± 1.9	82.20 ± 0.5	86.94 ± 1.8	77.39 ± 1.8	82.16 ± 0.5
RelationNet	83.47 ± 1.1	86.93 ± 2.9	79.26 ± 1.4	83.09 ± 1.5	81.58 ± 0.7	86.66 ± 1.4	76.59 ± 0.7	81.63 ± 0.8
Baseline	83.75 ± 1.0	89.81 ± 4.0	78.11 ± 2.8	83.96 ± 1.0	83.33 ± 0.4	86.93 ± 1.2	**79.76 ± 1.5**	83.35 ± 0.4
Baseline++	84.26 ± 1.2	90.22 ± 2.9	78.75 ± 4.8	84.18 ± 0.7	83.50 ± 0.3	88.56 ± 1.0	78.39 ± 1.6	83.47 ± 0.4
RFS	83.56 ± 1.0	90.29 ± 1.8	76.85 ± 4.0	83.57 ± 1.5	83.08 ± 0.4	88.39 ± 0.8	77.94 ± 1.4	83.16 ± 0.4
ANIL	83.64 ± 0.4	88.91 ± 1.3	78.87 ± 1.6	83.89 ± 0.4	83.70 ± 0.3	88.14 ± 0.9	79.53 ± 0.5	83.83 ± 0.4
RENet	84.03 ± 1.2	90.95 ± 2.6	77.16 ± 3.8	84.06 ± 0.9	83.41 ± 0.3	88.92 ± 0.8	77.89 ± 1.1	83.40 ± 0.4
IGCNN-FC	**85.55 ± 1.0**	**91.92 ± 5.5**	**79.42 ± 5.9**	**85.67 ± 1.0**	**84.50 ± 0.4**	**90.31 ± 1.6**	78.81 ± 1.8	**84.56 ± 0.4**

Table 2
The effect of three different losses in Eq. (8).

$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$	ACC	SPE	SEN	AUC
√			81.93 ± 1.7	89.27 ± 4.1	74.13 ± 5.1	81.70 ± 1.6
√	√		82.01 ± 2.1	91.14 ± 2.0	74.23 ± 5.5	82.18 ± 1.8
√	√	√	83.00 ± 1.4	91.57 ± 1.0	74.37 ± 2.8	82.94 ± 1.4

4.3. Experimental analysis

Table 1 presents the performance of different deep few-shot learning frameworks on Shenzhen and NIH datasets. It suggests that the proposed framework can obtain superior performance over the other competitors in almost all cases. For example, when 50% of the whole data are used for training on the Shenzhen dataset, the gain of the proposed IGCNN-FC is 1.53%, 1.88%, 0.85% and 1.77% over the best competitors in ACC, SPE, SEN and AUC, respectively. In summary, experimental results on two Chest X-rays datasets demonstrate that our method achieves superior performance over recent SOTA methods. In addition, our method has statistically significant difference from every comparison method because the p-values of all cases are < 0.038 on the paired-sample t-tests at the 95% significance level on disease diagnosis task. The possible reasons are: (1) The comparative methods might simply regard that all image patches have the same contribution, while our method can explore the significant mid- and high-level features by introducing the multi-scale loss-based attention. (2) Although some comparative methods also employ data augmentation to alleviate model overfitting, they cannot remove the inductive bias of data in source classes, because it might introduce uncertain correlation information between samples and classes during training (Liu, Fu, Xu, et al., 2021). By contrast, the proposed method mixes images and patches to capture more disentangled information, thereby alleviating the bias.

4.4. Ablation analysis

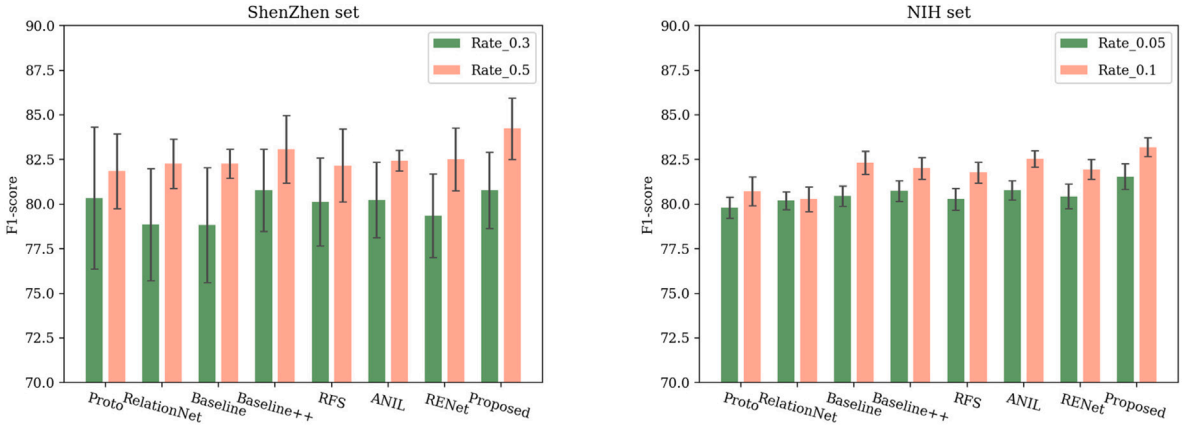
We conduct ablation experiments on the Shenzhen dataset (30%) to investigate the individual contribution corresponding to each component in the IGCNN-FC. Table 2 presents effect of three different losses in Eq. (8). Specifically, we progressively add different losses to our method. When each loss is added to IGCNN-FC, the corresponding performance is improved. It illustrates that both $\mathcal{L}_2$ and $\mathcal{L}_3$ have indispensable contributions. Additionally, to investigate the influence of multi-scale loss-based attention, we carry out experiments on three cases: (1) without employing any attention mechanism (w/o), (2) only using high-level features (high-level) and (3) using both mid- and high level features (multi-scale), and show their results in Table 3. It suggests that loss-based attention with high-level features can obtain superior performance over that without using any attention mechanism, while multi-scale loss-based attention with mid- and high-level features can attain better performance than that only using high-level features. Thus, multi-scale loss-based attention is beneficial to boosting the model performance.

4.5. F_1-Score

The F_1-score comprehensively considers the ACC and recall of the classification model. While increasing precision and recall as much as possible, it is also desirable that the difference between them be as slight as possible. Fig. 2 presents the cases of F_1-score

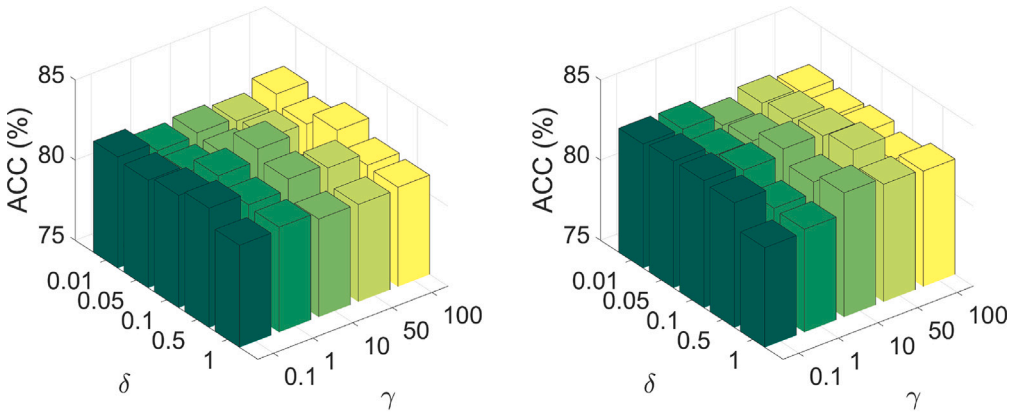
Table 3
The effect of multi-scale loss-based attention.

Attention	ACC	SPE	SEN	AUC
w/o	82.32 ± 2.7	91.22 ± 2.1	71.20 ± 5.1	82.21 ± 2.3
Attention-4	82.71 ± 1.6	91.36 ± 1.9	72.16 ± 3.9	82.66 ± 1.3
IGCNN-FC	83.00 ± 1.4	91.57 ± 1.0	74.37 ± 2.8	82.94 ± 1.4



(a) ShenZhen set (b) NIH set

Fig. 2. Comparison to prior work on the Shenzhen set and NIH set, respectively.



(a) ShenZhen set (b) NIH set

Fig. 3. The classification accuracy of IGCNN-FC at different parameter settings on δ and γ.

of IGCNN-FC on Shenzhen set and NIH set, respectively. We can find that the IGCNN-FC performs well overall, especially in the NIH set, where we outperform the contrast method at two different labeling rates. This further indicates that IGCNN-FC has a good performance in the medical image classification task.

5. Parameter analysis

We investigate the role of δ and γ in the IGCNN-FC. Specifically, we vary the values of δ and γ during the range of [0.01, 0.05, 0.1, 0.5, 1] and [0.1, 1, 10, 50, 100], respectively, to visualize the classification results of the two datasets in Fig. 3. As we can see, the proposed IGCNN-FC method can obtain the best or sub-best performance when $\delta \in [0.1, 1]$ and $\gamma \in [10, 100]$.

5.1. Interpretability

In terms of medical tasks, the interpretability of the deep learning model helps doctors understand the reasons for the decision-making given by the model and then make an accurate, efficient, and convincing diagnosis. We compare ResNet18+Grad-CAM with

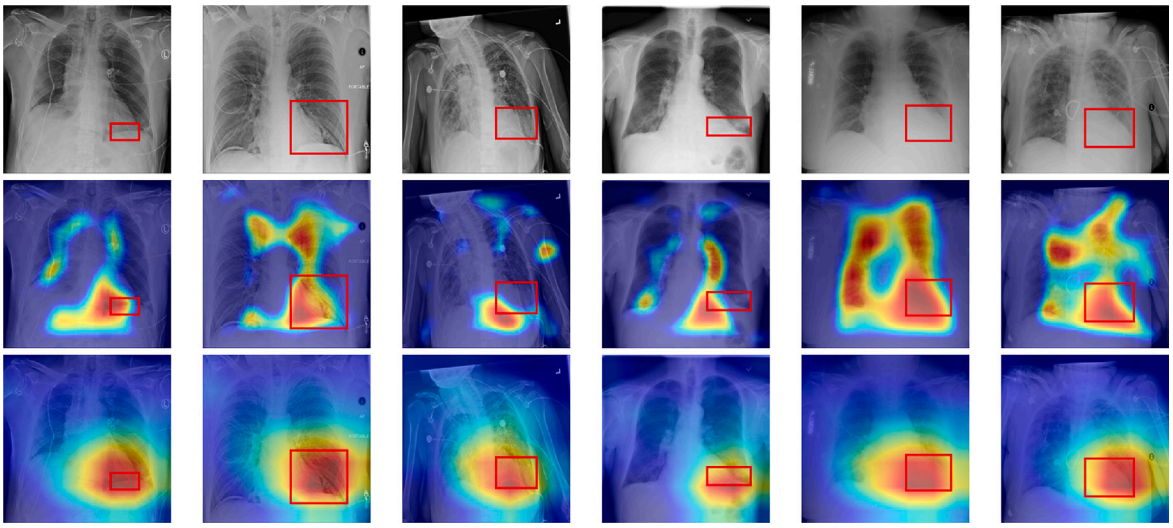


Fig. 4. Several heat maps of images from the NIH dataset by using ResNet18+Grad-CAM and IGCNN-FC. The first, second and third rows present the original image, heat maps generated by Grad-CAM and multi-scale loss-based attention, respectively.

IGCNN-FC, and present their heat maps in Fig. 4. It suggests that both Grad-CAM and multi-scale loss-based attention can interpret class-discriminative features. However, multi-scale loss-based attention can locate more accurate features. This might be because multi-scale loss-based attention can select significant features and meanwhile remove trivial features for image classification during model training.

6. Conclusion

In this paper, we propose a novel interpretable deep framework, IGCNN-FC, to mine the significant features for few Chest X-rays analysis. To fulfill this goal, we design multi-scale loss-based attention to explore and interpret the significance of mid- and high-level features, employ mixup to enlarge training data, and leverage self-distillation to further boost model's generalization capability. Experimental results on two Chest X-rays datasets demonstrate that our method can not only achieve superior performance over recent SOTA methods, but also boost model interpretability. Because we only consider the unimodal data in our experiments, and multi-modal data might provide better model performance and more comprehensive diagnosis opinions, in future work, we will combine Chest X-rays, clinical medical records (text) and biological (genome and proteome) data to build a cross-modal medical analysis framework to further improve the effectiveness of diagnosis and treatment.

CRediT authorship contribution statement

Mengmeng Zhan: Original idea, Methodology, Software, Writing – original draft. **Xiaoshuang Shi:** Methodology, Supervision. **Fangqi Liu:** Visualization, Writing – original draft. **Rongyao Hu:** Data curation, Writing – review & editing.

Data availability

Data will be made available on request.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Grant No. 61876046), Medico-Engineering Cooperation Funds from University of Electronic Science and Technology of China (No. ZYGX2022YGRH009 and ZYGX2022YGRH014), the Guangxi "Bagui" Teams for Innovation and Research, China. All authors read and approved the final manuscript.

References

Andonian, A., Chen, S., & Hamid, R. (2022). Robust cross-modal representation learning with progressive self-distillation. In *CVPR* (pp. 16430–16441).

Barnett, A. J., Schwartz, F. R., Tao, C., Chen, C., Ren, Y., Lo, J. Y., et al. (2021). A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence*, *3*(12), 1061–1070.

Beyer, L., Zhai, X., Royer, A., Markeeva, L., Anil, R., & Kolesnikov, A. (2022). Knowledge distillation: A good teacher is patient and consistent. In *CVPR* (pp. 10925–10934).

Bozorgtabar, B., Mahapatra, D., Vray, G., & Thiran, J. P. (2020). SALAD: Self-Supervised Aggregation Learning for Anomaly Detection on X-rays. In *MICCAI* (pp. 468–478). Springer.

Cai, T., Li, J., Mian, A. S., Sellis, T., Yu, J. X., et al. (2020). Target-aware holistic influence maximization in spatial social networks. *IEEE Transactions on Knowledge and Data Engineering*.

Candemir, S., Jaeger, S., Palaniappan, K., et al. (2013). Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Transactions on Medical Imaging*, *33*(2), 577–590.

Cao, A. Q., Puy, G., Boulch, A., & Marlet, R. (2021). PCAM: Product of Cross-Attention Matrices for rigid registration of point clouds. In *ICCV* (pp. 13229–13238).

Cen, J., Yun, P., Cai, J., Wang, M. Y., & Liu, M. (2021). Deep metric learning for open world semantic segmentation. In *ICCV* (pp. 15333–15342).

Chen, Z., Fu, Y., Wang, Y. X., Ma, L., Liu, W., & Hebert, M. (2019). Image deformation meta-networks for one-shot learning. In *CVPR* (pp. 8680–8689).

Chen, W. Y., Liu, Y. C., Kira, Z., et al. (2018). A closer look at few-shot classification. In *ICLR*.

Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F., & Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, *5*(6), 493–497.

Dabouei, A., Soleymani, S., Taherkhani, F., & Nasrabadi, N. M. (2021). Supermix: Supervising the mixing data augmentation. In *CVPR* (pp. 13794–13803).

Deng, C., Wang, M., Liu, L., Liu, Y., & Jiang, Y. (2021). Extended feature pyramid network for small object detection. *IEEE Transactions on Multimedia*.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML* (pp. 1126–1135). PMLR.

Gan, J., Hu, R., Mo, Y., Kang, Z., Peng, L., Zhu, Y., et al. (2022). Multigraph fusion for dynamic graph convolutional network. *IEEE Transactions on Neural Networks and Learning Systems*, http://dx.doi.org/10.1109/TNNLS.2022.3172588.

Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021). Knowledge distillation: A survey. *IJCV*, *129*(6), 1789–1819.

Guo, M. H., Xu, T.-X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., et al. (2022). Attention mechanisms in computer vision: A survey. *Computational Visual Media*, 1–38.

Hariharan, B., & Girshick, R. (2017). Low-shot visual recognition by shrinking and hallucinating features. In *ICCV* (pp. 3018–3027).

He, K., Girshick, R., & Dollár, P. (2019). Rethinking imagenet pre-training. In *ICCV* (pp. 4918–4927).

Hong, M., Choi, J., & Kim, G. (2021). Stylemix: Separating content and style for enhanced data augmentation. In *CVPR* (pp. 14862–14870).

Hou, R., Chang, H., Ma, B., Shan, S., & Chen, X. (2019). Cross attention network for few-shot classification. *Advances in Neural Information Processing Systems*, *32*.

Hu, H., Gu, J., Zhang, Z., Dai, J., & Wei, Y. (2018). Relation networks for object detection. In *CVPR* (pp. 3588–3597).

Huisman, M., Van Rijn, J. N., & Plaat, A. (2021). A survey of deep meta-learning. *Artificial Intelligence Review*, *54*(6), 4483–4541.

Jaeger, S., Karargyris, A., Candemir, S., et al. (2013). Automatic tuberculosis screening using chest radiographs. *IEEE Transactions on Medical Imaging*, *33*(2), 233–245.

Kaissis, G., Ziller, A., Passerat-Palmbach, J., et al. (2021). End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nature Machine Intelligence*, *3*(6), 473–484.

Kang, D., Kwon, H., Min, J., & Cho, M. (2021). Relational embedding for few-shot classification. In *ICCV* (pp. 8822–8833).

Kang, Z., Zhang, P., Zhang, X., Sun, J., & Zheng, N. (2021). Instance-conditional knowledge distillation for object detection. *Advances in Neural Information Processing Systems*, *34*, 16468–16480.

Kim, S., Kim, D., Cho, M., & Kwak, S. (2021). Embedding transfer with label relaxation for improved metric learning. In *CVPR* (pp. 3967–3976).

Koch, G., Zemel, R., Salakhutdinov, R., et al. (2015). Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop, vol. 2*. Lille.

Li, Z., Zhou, F., Chen, F., & Li, H. (2017). Meta-sgd: Learning to learn quickly for few-shot learning. arXiv preprint arXiv:1707.09835.

Liu, C., Fu, Y., Xu, C., et al. (2021). Learning a few-shot embedding model with contrastive learning. In *AAAI, vol. 35* (pp. 8635–8643).

Lu, D., Luo, Q., Chen, R., Zhuansun, Y., et al. (2020). Chemical multi-fingerprinting of exogenous ultrafine particles in human serum and pleural effusion. *Nature Communications*, *11*(1), 1–8.

Ma, J., Li, D., Zhu, H., Li, C., Zhang, Q., & Qiao, Y. (2022). GAFM: A knowledge graph completion method based on graph attention faded mechanism. *Information Processing & Management*, *59*(5), Article 103004.

Mo, Y., Peng, L., Xu, J., Shi, X., & Zhu, X. (2022). *Simple unsupervised graph representation learning*. AAAI.

Nguyen, T., Luu, T., Pham, T., Rakhimkul, S., & Yoo, C. D. (2021). Robust MAML: Prioritization task buffer with adaptive learning process for model-agnostic meta-learning. In *ICASSP* (pp. 3460–3464).

Noguchi, S., Nishio, M., Yakami, M., Nakagomi, K., & Togashi, K. (2020). Bone segmentation on whole-body CT using convolutional neural network with novel data augmentation techniques. *Computers in Biology and Medicine*, *121*, Article 103767.

Osahor, U., & Nasrabadi, N. M. (2022). Ortho-shot: Low displacement rank regularization with data augmentation for few-shot learning. In *WACV* (pp. 2200–2209).

Peng, L., Hu, R., Kong, F., Gan, J., Mo, Y., Shi, X., et al. (2022). Reverse graph learning for graph neural network. *IEEE Transactions on Neural Networks and Learning Systems*, http://dx.doi.org/10.1109/TNNLS.2022.3161030.

Qiao, S., Chen, L. C., & Yuille, A. (2021). Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In *CVPR* (pp. 10213–10224).

Raghu, A., Raghu, M., Bengio, S., & Vinyals, O. (2019). Rapid learning or feature reuse? Towards understanding the effectiveness of MAML. In *ICLR*.

Shang, Y., Duan, B., Zong, Z., Nie, L., & Yan, Y. (2021). Lipschitz continuity guided knowledge distillation. In *ICCV* (pp. 10675–10684).

Shang, Z., Xie, H., Zha, Z., Yu, L., Li, Y., & Zhang, Y. (2021). PRRNet: Pixel-Region Relation Network for face forgery detection. *Pattern Recognition*, *116*, Article 107950.

Shi, X., Xing, F., Xie, Y., Zhang, Z., Cui, L., & Yang, L. (2020). Loss-based attention for deep multiple instance learning. *AAAI*, *34*(04), 5742–5749.

Shi, X., Xing, F., Xu, K., Chen, P., Liang, Y., Lu, Z., et al. (2020). Loss-based attention for interpreting image-level prediction of convolutional neural networks. *IEEE Transactions on Image Processing*, *30*, 1662–1675.

Shu, Y., Cao, Z., Wang, C., Wang, J., & Long, M. (2021). Open domain generalization with domain-augmented meta-learning. In *CVPR* (pp. 9624–9633).

Silva, W., Poellinger, A., Cardoso, J. S., & Reyes, M. (2020). Interpretability-guided content-based medical image retrieval. In *MICCAI* (pp. 305–314). Springer.

Singh, R., Hie, B. L., Narayan, A., & Berger, B. (2021). Schema: metric learning enables interpretable synthesis of heterogeneous single-cell modalities. *Genome Biology*, *22*(1), 1–24.

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems*, *30*.

Song, X., Li, J., Lei, Q., Zhao, W., Chen, Y., & Mian, A. (2022). Bi-CLKT: Bi-graph Contrastive Learning Based Knowledge Tracing. *Knowledge-Based Systems*, *241*, Article 108274.

Sterling, T. R., Njie, G., Zenner, D., et al. (2020). Guidelines for the treatment of latent tuberculosis infection: recommendations from the National Tuberculosis Controllers Association and CDC, 2020. *American Journal of Transplantation*, *20*(4), 1196–1206.

Sun, Y., DeJaco, R. F., Li, Z., Tang, D., Glante, S., Sholl, D. S., et al. (2021). Fingerprinting diverse nanoporous materials for optimal hydrogen storage conditions using meta-learning. *Science Advances*, *7*(30), eabg3983.

Sung, F., Yang, Y., Zhang, L., et al. (2018). Learning to compare: Relation network for few-shot learning. In *CVPR* (pp. 1199–1208).

Thulasidasan, S., Chennupati, G., Bilmes, J. A., Bhattacharya, T., & Michalak, S. (2019). On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, *32*.

Tian, Y., Wang, Y., Krishnan, D., Tenenbaum, J. B., & Isola, P. (2020). Rethinking few-shot image classification: a good embedding is all you need? In *ECCV* (pp. 266–282). Springer.

Vinyals, O., Blundell, C., Lillicrap, T., et al. (2016). Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, *29*.

Wang, X., Peng, Y., Lu, L., et al. (2017). ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR* (pp. 3462–3471).

Wang, L., & Yoon, K. J. (2021). Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, Y., Zhang, J., Kan, M., Shan, S., & Chen, X. (2020). Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *CVPR* (pp. 12275–12284).

Wu, Y., Wang, H., & Wu, F. (2017). Automatic classification of pulmonary tuberculosis and sarcoidosis based on random forest. In *2017 10th international congress on image and signal processing, biomedical engineering and informatics* (pp. 1–5). IEEE.

Xu, J., Ren, Y., Tang, H., Yang, Z., Pan, L., Yang, Y., et al. (2022). Self-supervised discriminative feature learning for deep multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, http://dx.doi.org/10.1109/TKDE.2022.3193569.

Xue, G., Zhong, M., Li, J., Chen, J., Zhai, C., & Kong, R. (2022). Dynamic network embedding survey. *Neurocomputing*, *472*, 212–223.

Yoon, J., Kang, D., & Cho, M. (2022). Semi-supervised domain adaptation via sample-to-sample self-distillation. In *ICCV* (pp. 1978–1987).

Yuan, C., Zhong, Z., Lei, C., Zhu, X., & Hu, R. (2021). Adaptive reverse graph learning for robust subspace learning. *Information Processing & Management*, *58*(6), Article 102733.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV* (pp. 6023–6032).

Zhang, Q., Hann, E., Werys, K., Wu, C., Popescu, I., Lukaschuk, E., et al. (2020). Deep learning with attention supervision for automated motion artefact detection in quality control of cardiac T1-mapping. *Artificial Intelligence in Medicine*, *110*, Article 101955.

Zhang, K., & Zhuang, X. (2022). CycleMix: A holistic strategy for medical image segmentation from scribble supervision. In *CVPR* (pp. 11656–11665).

Zhu, Y., Ma, J., Yuan, C., & Zhu, X. (2022). Interpretable learning based dynamic graph convolutional networks for alzheimer's disease analysis. *Information Fusion*, *77*, 53–61.

Zhu, X., Song, B., Shi, F., et al. (2021). Joint prediction and time estimation of COVID-19 developing severe symptoms using chest CT scan. *Medical Image Analysis*, *67*, Article 101824.

Zhu, X., Zhang, S., Zhu, Y., Zhu, P., & Gao, Y. (2020). Unsupervised spectral feature selection with dynamic hyper-graph learning. *IEEE Transactions on Knowledge and Data Engineering*, http://dx.doi.org/10.1109/TKDE.2020.3017250.