Neural Processing Letters (2023) 55:2175–2191
https://doi.org/10.1007/s11063-022-10940-4

A Novel Distant Domain Transfer Learning Framework for Thyroid Image Classification

Fenghe Tang[1] · Jianrui Ding[1] · Lingtao Wang[1] · Chunping Ning[2]

Accepted: 16 June 2022 / Published online: 25 June 2022

Abstract
Medical ultrasound imaging technology is currently the preferred method for early diagnosis of thyroid nodules. Radiologists' analysis of ultrasound images is highly dependent on their clinical experience and is susceptible to intra- and inter-observer variability. Although end-to-end deep learning technique can address these limitations, the difficulty of acquiring annotated medical image makes it very challenging. Transfer learning can alleviate the problems, but the large gap between source and target domain will lead to negative transfer. In this paper, a novel transfer learning method with distant domain high-level feature fusion (DHFF) model is proposed. It reduces the distribution distance between the source domain and the target domain while maintaining the characteristics of respective domains, which can avoid excessive feature fusion while enabling the model to learn more valuable transfer knowledge. The DHFF is validated by multiple public source and private target datasets in experiments. The results show that the classification accuracy of DHFF is up to 88.92% with thyroid ultrasound auxiliary source domains, which is up to 8% higher than existing transfer and distant transfer algorithms.

Keywords Transfer learning · Thyroid image classification · Distant domain · Feature Fusion

✉ Jianrui Ding
jrding@hit.edu.cn

Fenghe Tang
21s130300@stu.hit.edu.cn

Lingtao Wang
hit_wanglingtao@163.com

Chunping Ning
152081340@qq.com

[1] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

[2] Ultrasound Department, The Affiliated Hospital of Qingdao University, Qingdao, China

1 Introduction

Ultrasound imaging technology is a non-invasive, non-radiation, low cost and real time detection method. It has widely used in the detection of thyroid, fetal, mammary gland and gonadal tissue [1]. However, due to its low contrast, Manual image analysis is time-consuming and laborious. At the same time, it will be affected by subjective factors such as radiologists' experience and mental state, which is prone to misdiagnosis. Automatic medical image analysis techniques can effectively overcome the above limitations. Generally, it can be roughly classified into two main categories: hand-crafted feature-based methods and the data-driven methods. The pipeline of hand-craft feature methods frequently involves feature extraction and classification. Despite their rapid development in recent years, handcrafted features are highly dependent on expert knowledge. Moreover, handcrafted features in some sense ignore high-level abstract information that is not visible to the human eye in ultrasound images and it can only exploit the low-level information, such as image texture [2, 3], geometry morphology [4], and statistical distributions [5]. Such methods usually require further employ classifiers to conduct classification. Therefore, only given the highly discriminative features, this method can solve the recognition problem well. On the contrary, data-driven methods, without the need of hand-crafted feature description, can greatly improve the classification performance of medical images by using the convolution neural networks (CNNs) [6]. At present, deep learning technology in data-driven approaches has become the mainstream method of image analysis and understanding [7]. It depends on large-scale labeled training data. However, compared with natural scene images, the collection and annotation of medical images is difficult and expensive, which brings great challenges to the application of deep learning technology in medical domain. Transfer learning [8] is one of the effective methods to solve small data learning, which has been widely used in medical image analysis [9]. Theoretically, it attempts to build a robust model by transferring the knowledge learned from the source domain with large-scale training data to the target domain with a small amount of data. However, low or even irrelevance between the source and target domains may lead to negative transfer [8], which causes the knowledge generated by the source domain negatively affects the target domain. How to transfer the knowledge beneficial to medical image analysis from the source domain is a challenging problem. In this paper, we propose a method to transfer valuable knowledge from distant domains which have low correlation or even seemingly unrelated with target domain, and apply it to thyroid ultrasound image classification.

Distant domain transfer learning (DDTL) [10] is a new transfer learning method. It is inspired by the human ability to learn new things by integrating knowledge obtained from several seemingly independent things. Considering the variation between distant domain and target domain, the auxiliary domain is served as an intermediary bridge to narrow the gap between them. Most of the existing methods [10–13] apply simple auto-encoders as feature extractors. The extracted features are more representative of the low-level details of the image and lack the ability to represent the domain Therefore they are greatly affected by the variation among domains. The performance and stability are unsatisfactory. The existing distant domain transfer learning methods in the past are summarized as listed in Table 1 and we introduce the limitation of these methods. Inspired by feature-based [12] and instance-based [11] DDTL methods, we propose a DDTL method which extracts high-level semantic features from different domains and performs distant domain feature fusion. An auxiliary domain is also adopted as a bridge to reduce the gap between source domain and target domain. It can transfer valuable knowledge to the target domain, reduce the cost and effort of collecting training data in target domain, and suppress the negative effects resulting from

Table 1 Summary of distant domain transfer learning method

DDTL Techniques	Method	Limitations
TTL [10]	Instances-based	Highly dependent intermediate domain, which is selected by users manually
SLA [11]	Instances-based	Need to adjust the conditions for selecting instances according to different tasks
Xie et al. [21]	Feature-based	Requires a large amount of labeled intermediate data, which can be too expensive to apply
DFF [12]	Feature-based	Knowledge transfer only for low-level semantic features
AM-DDTL [13]	Feature-based	The feature extraction is computationally expensive

the irrelevant parts of distant domains. The flow chart of our method is shown in Fig. 1. Firstly, crop region of interest in thyroid ultrasound dataset and augment the source and the target domain dataset. Secondly, source and target domain high-level semantic features are extracted by high-level semantic features extractor (CNN). And then utilize an auto encoder-decoder to perform high-level semantic feature fusion, and the decoder is applied to maintain the diversity of each domain. Finally, a target classifier is deployed for target domain classification. The effectiveness of this method is verified on thyroid ultrasound dataset. Our main contributions are summarized as follows:

- The feature extraction network is redesigned to mine the information transferred from the source domain to the target domain that is beneficial to the target task. Based on resnet50, a lightweight feature extraction network is designed to extract high-level semantic features that can better represent the source and target domains.
- Propose a novel distant domain high-level semantic feature fusion method. A high-level semantic feature adaption encoder with domain distance measure is designed to discover valuable knowledge across different domains.
- To improve the generalization performance of transfer learning, the diversity of source and target domains is preserved while narrowing the difference between source and target domains. A decoder and content loss are added to narrow the content gap between the reconstructed features and the input features, which helps maintain the invariance of the source domain and the target domain.
- The effectiveness of the proposed method is verified by extensive experiments on open-source datasets, thyroid and breast ultrasound images.

The rest of the paper is organized as follows. In Sect. 2 and Sect. 3, we briefly describe the recent work of DDTL and formulate the DDTL problem definition. In Sect. 4 we present the details of the proposed method. In Sect. 5, we present experimental results and analysis. Finally, in Sect. 6, we conclude our work.

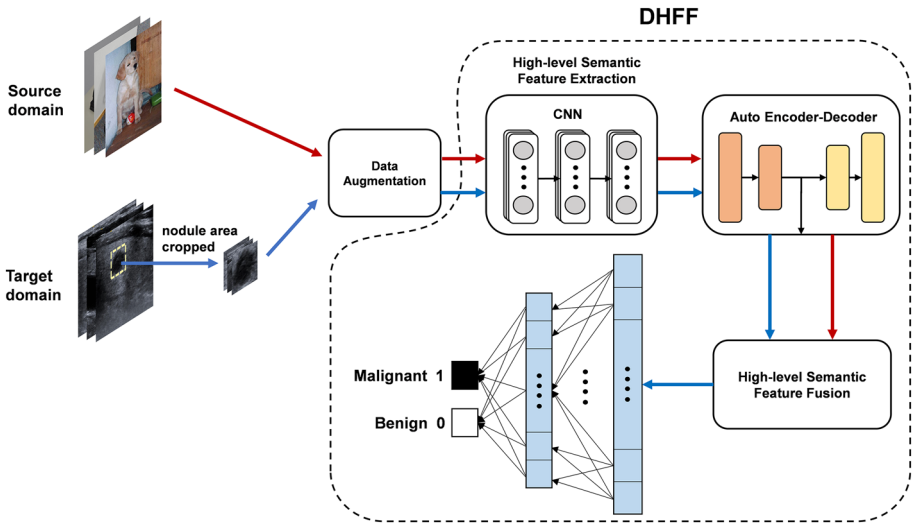


Fig. 1 The flow chart of our method

2 Related Work

Deep learning models, especially convolutional neural networks (CNN), have shown advantages over other traditional learning methods in thyroid ultrasound image tasks [6, 14]. Many scholars use CNN to achieve the task of thyroid nodule image classification and segmentation [15–17]. Unfortunately, the lack of annotated training data limits the performance of most models. The transfer learning method can effectively solve the above problems, and has been applied to thyroid ultrasound image task [14, 18, 19].

However, traditional transfer learning methods are still limited by the distribution variation between the source and target domains. Recent studies [20] have shown that fine-tuning a pre-trained model on natural images may not help to significantly improve the accuracy of medical image classification. Unlike fine-tuning, DDTL focuses on making use of the knowledge in distant domains to improve the performance of the task in the target domain. Moreover, in the previous research of DDTL algorithm, Tan [10, 11] introduced two instance-based algorithms: Transitive Transfer learning (TTL) and Selective Learning Algorithm (SLA). TTL transfers knowledge between text data in the source domain and the image data in the target domain by using annotated image data as a bridge. SLA select helpful instances from several unrelated auxiliary domains to expand the source domain's volume. However, the limitations of these two methods are that they are unstable and can only be used for binary classification. Xie [21] proposed a feature-based method to predict poverty using satellite imagery. This method exploits nighttime light intensity as an intermediate domain to transfer the knowledge learned from natural images during the day to high-resolution satellite images, but this method requires a large amount of annotated data in auxiliary domain. Niu [12] proposed another feature-based method (DFF) to the classification and diagnosis of COVID-19 images. DFF does not require labeled data in source and intermediate domains, it can solve multi-classification problems as well. However, DFF cannot perform high-level semantic features fusion, and its model is only used to process low-noise lung CT images. Qin [13] added an attention mechanism into DDTL for extraction more effective information from

satellite images, however the processing of this method feature extraction is computationally expensive and it still unable to extract sufficient high-level semantic features.

Recent work [20] shows that low-level image features are more likely to be reused in transfer learning. However, these low-level image features in different domains are very similar, which cannot provide additional valuable domain knowledge to be transferred. On the contrary, high-level semantic features can better represent the essential characteristics of the domain. Therefore, extracting and fusing the high-level semantic features of the domain can maximize the learning of additional valuable knowledge in the distant domain.

In this paper, we aim to implement efficient transfer of high-level semantic information in distant domain through DDTL. The layer that can extract high-level semantic features is retained in CNN model, and a lightweight network is designed to realize efficient high-level semantic feature fusion between distant domain and target domain. Then an encoder-decoder network is utilized to retain the high-level features of each domain, increase the diversity and improve the generalization of the model. Finally, multiple loss functions are integrated to make a trade-off among classification, transferring knowledge and preserving the diversity.

3 Preliminary

3.1 Problem Definition

The goal of DDTL is to transfer the valuable information in distant source domains to target domain. An auxiliary domain is also used as a bridge between the source domain and the target domain.

According to the DDTL problem, we assume that a small amount of target domain data is not enough to train a robust model. The target domains T with fewer annotated data are denoted as:

$$T = [(x_1^1, y_1^1), ..., (x_1^{n_{T_1}}, y_1^{n_{T_1}})], ..., [(x_{T_N}^1, y_{T_N}^1), ..., (x_{T_N}^{n_{T_N}}, y_{T_N}^{n_{T_N}})] \tag{1}$$

where n_{T_i} and T_N represent the number of samples in ith target domain and the number of target domains, x_i^j is the jth sample of the ith target domain, y_i^j is corresponding label. Then we denote the unlabeled source domain S as:

$$S = \left\{(x_1^1, ..., x_1^{n_{S_1}}), ..., (x_{S_N}^1, ..., x_{S_N}^{n_{S_N}})\right\} \tag{2}$$

where n_{S_i} and S_N represent the number of samples in ith source domain and the number of source domains, furthermore, we denote the unlabeled auxiliary domain A as:

$$A = \left\{(x_1^1, ..., x_1^{n_{A_1}}), ..., (x_{A_N}^1, ..., x_{A_N}^{n_{A_N}})\right\} \tag{3}$$

where n_{A_i} and A_N represent the number of samples in ith auxiliary domain and the number of auxiliary domains. Let the marginal distribution and conditional distribution of source domain data be denoted as $p_S(x)$,$p_S(y|x)$ the distribution of target domain data be denoted as $p_T(x)$,$p_T(y|x)$, and the distribution of auxiliary domain be denoted as $p_A(x)$, $p_A(y|x)$, we have the following assumptions:

$$p_S(x) \neq p_T(x), p_T(x) \neq p_A(x) \tag{4}$$

$$p_{T_1}(y|x) \neq p_{T_2}(y|x) \neq ... \neq p_{T_N}(y|x) \tag{5}$$

3.2 Maximum Mean Discrepancy

Maximum Mean Discrepancy (MMD) [22] is an important statistical indicator to measure the discrepancy between the source and target domain distributions ($p_S(x)$ and $p_T(x)$). Given the distributions s and t over two domains, MMD is presented as:

$$MMD(s,t) = \sup_{\|\varphi\|_{\mathcal{H}} \leq 1} \|E_{X^s \sim S}[\varphi(X^s)] - E_{X^t \sim s}[\varphi(X^t)]\|_{\mathcal{H}} \tag{6}$$

where φ represents the kernel function that maps the original data to the Reproduced Kernel Hilbert Space (RKHS) [22]. The empirical estimate of MMD is defined as:

$$MMD(S,T) = \left\| \frac{1}{M} \sum_{i}^{M} \varphi(s_i) + \frac{1}{N} \sum_{j}^{N} \varphi(t_j) \right\|_H \tag{7}$$

where M and N are the number of instances in the source and target domains.

4 Distant High-Level Feature Fusion Model (DHFF)

In this section, we present the proposed DHFF. Firstly, we introduce the redesigned high-level semantic feature extractor. After that, we present the implementation details of the encoder-decoder network and define the content loss for maintaining different domains invariance. And then, we denote high-level semantic feature adaption methods. Finally, the classification loss and the overall loss of DHFF are defined.

Our proposed model DHFF is shown in Fig. 2. It mainly consists of four parts: High-level Semantic Feature Extraction, High-level Semantic Feature Encoder-Decoder, High-level Semantic Feature Adaption and Target domain Classification. The last three parts correspond to three types of losses: content loss, domain loss, and classification loss.

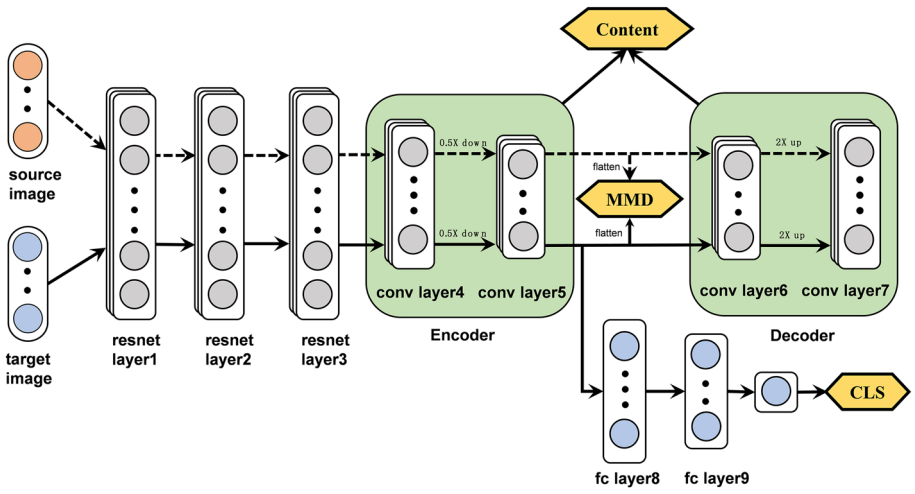


Fig. 2 DHFF model

4.1 High-level Semantic Feature Extraction

Inspired by a recent study [20], we use hybrid approaches to transfer learning where a subset of pretrained weights are used, and the high-level parts of the network are redesigned and made more lightweight. In this paper, we design a lightweight feature extraction network based on ResNet50 [23]. In the structure of ResNet50, the closer to the output, the more the features can reflect the high-level semantic information of the input, which is more conducive to the classification task, and the greater the correlation with the domain. In DDTL, we want to transfer features from the source domain to the target domain to improve the performance of tasks in the target domain. Therefore, the correlation between the transferred features and the source domain should be reduced. In the structure of ResNet50, the features closer to the input can reflect the low-level features of the image, such as edge, texture and other information. These features cannot represent the domain well. As a trade-off, we retain the first three layers of ResNet50, extract the information that can both reflect the source domain and target domain. The redesigned feature extractor can also prevent the large fluctuation of weight when trained on different domain data and accelerate the convergence.

4.2 Encoder-Decoder Network and Content Loss

In order to maintain the diversity of features, avoid excessively narrowing the inter-domain distribution distance. In [24], it exploits autoencoder pair to minimize the reconstruction error on all the training instances. Autoencoder is an unsupervised feedforward neural network capable of efficient feature extraction and dimensionality reduction. However, using the it directly may not be able to effectively learn valuable high-level semantic information, and its performance is unstable. Therefore, an encoder-decoder network is applied to our study to extract and preserve inter-domain high-level semantic features. Generally, a convolutional autoencoder pair includes an input layer, an output layer, an up-sampling layer and multiple convolutional layers. The output from feature extractor in 4.1 is encoded with encoder and decoded by the decoder to reconstruct the high-level semantic features. The specific structure of the encoder-decoder is shown in Fig. 3. The encoder includes two pooling layers and two convolutional layers, conv layer4 and conv layer5 respectively. Specifically, a 3x3 convolution kernel with pad of 1 and stride of 1 is applied, and a 2x2 max pooling layer is used for down-sampling. The decoder consists of two convolutional layers and two up-sampling layers. We adopt 2x2 for up-sampling to maintain the same quality of reconstructed feature maps. The standard process of encode-decode can be demonstrated as:

$$Encoding: f_{abstract} = Encoder(f), Decoding: \hat{f} = Decoder(f_{abstract}) \quad (8)$$

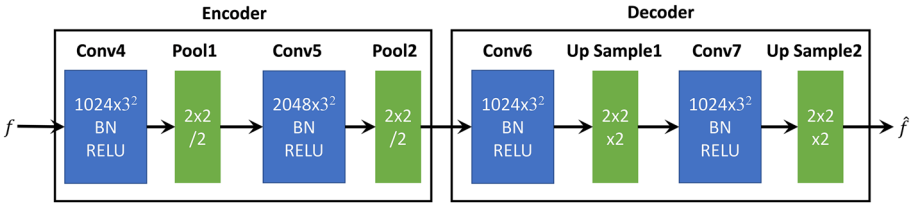


Fig. 3 The specific structure of the encoder-decoder

where f is the output of redesigned ResNet50 model, $f_{abstract}$ is a higher-level abstract feature of f, $\hat{f}$ is the high-level semantic feature after reconstruction. Our motivation is that, if the knowledge transferred from the source domain is helpful to the target domain, the encoder and decoder can be learned to maintain the integrity of the respective domain features while minimizing the difference between the high-level semantic features of the source and target domains. The auxiliary domain is also used to prevent excessive fusion of high-level semantic features between different domains. Inspired by style transfer, we adopt content loss to measure the difference between reconstruction features and input features. It is defined as:

$$L_{content} = \frac{1}{n_S}\sum_{i=1}^{n_S}(\hat{f}_S^i - f_S^i) + \frac{1}{n_T}\sum_{j=1}^{n_T}(\hat{f}_T^j - f_T^j) \tag{9}$$

when adding a auxiliary domain, the content loss objective function to be minimized is formulated as:

$$L_{content} = \frac{1}{n_S}\sum_{i=1}^{n_S}(\hat{f}_S^i - f_S^i) + \frac{1}{n_T}\sum_{j=1}^{n_T}(\hat{f}_T^j - f_T^j) + \frac{1}{n_A}\sum_{k=1}^{n_A}(\hat{f}_A^k - f_A^k) \tag{10}$$

where f_S^i,f_T^j and f_A^k represent the high-level semantic features of a sample in source domain, target domains and auxiliary domains respectively, $\hat{f}_S^i$,$\hat{f}_T^j$ and $\hat{f}_A^k$ are the reconstructed features of the sample in the corresponding domains, n_S,n_T,n_A are the number of samples in source domain, target domain and auxiliary domain.

4.3 High-Level Semantic Feature Adaption

Minimizing the content loss can preserve the integrity of the respective domain features and ensure the diversity. However, in order to use the knowledge transferred from distant domain to help the target domain task, it is necessary to narrow the distance from the distant domain to the target domain. We introduce the Maximum Mean Discrepancy (MMD) mentioned in 3.2 as domain loss to measure the distribution distance between distant domain and target domain. It is defined as:

$$L_{domain} = MMD\left(\frac{1}{n_S}\sum_{i=1}^{n_S} f_{S_{abstract}}^i, \frac{1}{n_T}\sum_{j=1}^{n_T} f_{T_{abstract}}^j\right) \tag{11}$$

when adding auxiliary domains, it is formulated as:

$$\begin{aligned} L_{domain} =& MMD\left(\frac{1}{n_S}\sum_{i=1}^{n_S} f_{S_{abstract}}^i, \frac{1}{n_A}\sum_{k=1}^{n_A} f_{A_{abstract}}^k\right) \\ &+ MMD\left(\frac{1}{n_T}\sum_{j=1}^{n_T} f_{T_{abstract}}^j, \frac{1}{n_A}\sum_{k=1}^{n_A} f_{A_{abstract}}^k\right) \end{aligned} \tag{12}$$

where $f_{S_{abstract}}^i$,$f_{T_{abstract}}^j$ and $f_{A_{abstract}}^k$ represent the high-level semantic features extracted by encoder in 4.2.

4.4 Target Domain Classification

Two fully-connected layers are added after the encoder in Fig. 2 to build a target classifier. They can find the best combination of high-level semantic abstraction features for the target task. The cross-entropy loss $L_{classification}$ is adopted as classification loss. It is defined as follows:

$$L_{classification} = \frac{1}{n_T} \sum_{i=1}^{n_T} -(y_i log(p_T^i) + (1 - y_i) log(1 - p_T^i)) \tag{13}$$

where p_T^i is the prediction result of sample ith in target domain, y_i is the label corresponding to the sample ith.

Finally, the overall loss of DHFF can be expressed as:

$$\underset{\theta_F,\theta_E,\theta_D,\theta_C}{Minimize} \; L = L_{classification} + L_{content} + L_{domain} \tag{14}$$

where θ_F,θ_E,θ_D,θ_C are the parameters of the high-level semantic feature extractor, encoder, decoder and the classifier. Finally, all parameters in the network are optimized by minimizing the objective function L.

5 Experimentation

5.1 Datasets

As shown in Table 2, We evaluate the DHFF model with five datasets, three of which are open source datasets: Catech-256 [25], Office31 [26] and Stanford Open Source Thyroid Nodule Ultrasound Image Dataset (Thyroid Ultrasound Cine-clip) [27], private datasets: thyroid and breast ultrasound datasets acquired from different ultrasound machines at the Affiliated Hospital of Qingdao University. Catech-256 is a natural image dataset that includes labeled data of 256 different class. Then, Office-31 includes 31 different common office items and is composed of a collection of three different domains: “Webcam”, “Dslr” and “Amazon”. Furthermore, the Stanford open-source thyroid nodule dataset consists of 17,412 thyroid ultrasound images provided by 167 patients. Moreover, the thyroid ultrasound dataset provided by Affiliated Hospital of Qingdao University consists of 3,003 images collected by different ultrasound machines, 2268 malignant samples and 735 benign samples. And the breast ultrasound dataset consists of 725 images.

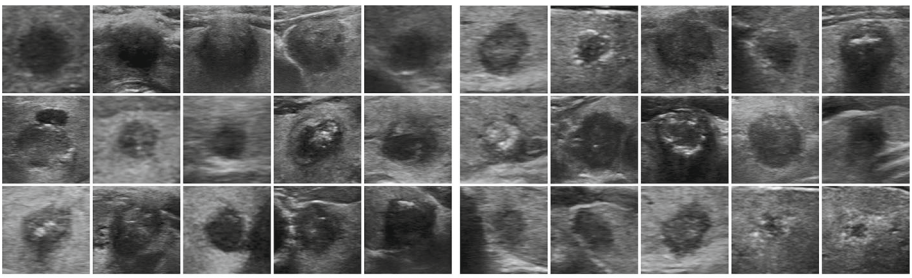
Some ultrasound thyroid images are shown in Fig. 4. Fig. 4a shows the benign nodules, and most of them have irregular shapes, smooth regions, and boundaries. Fig. 4b shows the malignant nodules, and most of them have irregular shapes, coarse regions, and boundaries. As can be seen from Fig. 4, there is some overlap in the image characteristics between benign and malignant thyroid nodule. Therefore, it is difficult to differentiate thyroid nodules based on low-level features such as gray scale and shape.

5.2 Baseline Model and Experiment Setups

As shown in Fig. 5, the modified ResNet50 model (ResNet50-baseline) is used as a baseline to verify the rationality and effectiveness of our proposed method (DHFF). It exploits ResNet50 as a feature extractor and implements domain-adaptive transfer learning by minimizing the marginal distribution of the two domains.

Table 2 dataset

Data Set	Classes num	Samples num	Label	Mask
Catech-256	256	30670	Yes	No
Office31	31	4110	Yes	No
Thyroid Ultrasound Cine-clip	2	17412	Yes	Yes
Thyroid Ultrasound	2	3003	Yes	Yes
Breast Ultrasound	2	725	Yes	Yes



(a) Benign **(b)** Malignant

Fig. 4 Illustration of thyroid nodules: (a) Benign nodules; (b) Malignant nodules

In single distant source domain transfer learning category, we choose six transfer learning algorithms including Deep Transfer Learning (DTL) [28], Co-Tuning [29], Selective Learning Algorithm (SLA) [11], feature adaption baseline model (ResNet50-baseline), Distant Feature Fusion (DFF) [12] and AM-DDTL [13]. The ResNet50-baseline and DHFF are initialized with parameters pretrained on the ImageNet dataset [30], while the other methods are trained from scratch respectively. For DTL, we first train a deep model in the source domain, and then train another deep model for the target domain by reusing the first several layers of the source model. And for Co-Tuning, we use ResNet50 as a source domain pretrain model. For SLA, DFF, DHFF, ResNet50-baseline and AM-DDTL, we use all the source domain, target domain data to learn a model. Meanwhile, in multiple source domains transfer learning category, we choose four transfer learning algorithms including SLA, ResNet50-baseline, DFF and AM-DDTL. We use all the source domain, target domain and auxiliary source domains data to learn a model.

In all experiments, as shown in Fig. 6, according to the ground truth boundary of the thyroid nodule delineated by the radiologist, the minimum circumscribed rectangle containing the nodule area is cropped from the image as the input. Meanwhile, in experiments 1-3, we use the full thyroid ultrasound dataset for experiments. In experiment 4 we conduct experiments with a small thyroid ultrasound dataset, which has 716 images. Specifically, it consists of 360 malignant samples and 356 benign samples. Besides, in our 1-3 experiments, the thyroid dataset is split by 6:2:2 for training, validating and testing respectively. Meanwhile, in our 4 experiments, the thyroid dataset is split by 7:2:1 for training, validating and testing respectively. Moreover, the random cropping and horizontal flipping are used for natural images, and horizontal flipping is used for medical ultrasound images data augmentation. Furthermore, each experiment was run eight times and the accuracy was averaged to remove performance fluctuations due to parameter initialization.

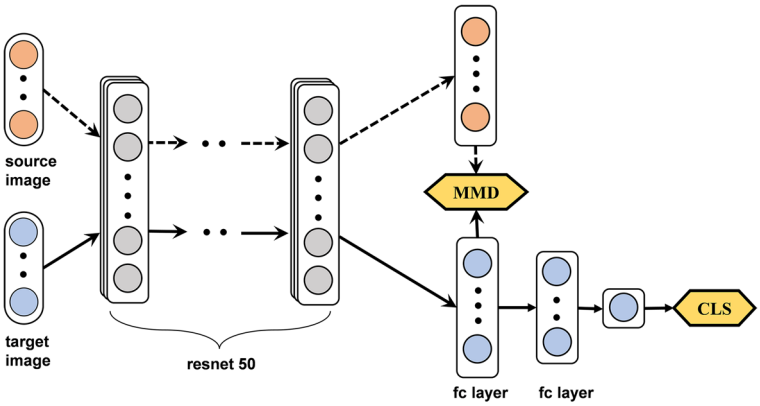


Fig. 5 baseline model based on Resnet50

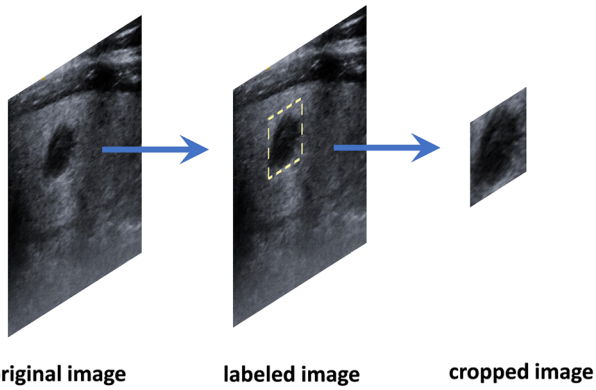


Fig. 6 thyroid ultrasound datasets preprocessing

In Experiment 1, six datasets, Catech-256, Amazon, Dslr, Webcam, Thyroid Ultrasound Cine-clip, and breast ultrasound datasets, are served as unlabeled source domains to explore the correlation of natural scene images and ultrasound images with targets domains. The experimental results are shown in Table 3. Then in Experiment 2, four datasets, Catech-256, Amazon, Dslr, Webcam are used as unlabeled source domains and breast ultrasound images is served as auxiliary domain, which can be regarded as related to target domain (thyroid ultrasound images). This experiment is utilized to verify whether the auxiliary domain close to the target domain is helpful for distant domain transfer. The experimental results are shown in Table 4. Finally, in Experiment 3, five datasets, Catech-256, Amazon, Dslr, Webcam, and breast ultrasound datasets are used as unlabeled source domain and Stanford thyroid ultrasound datasets is served as auxiliary domain. The experimental results are shown in Table 5.

5.3 Experimental Result Analysis

For single-source domain experiment, as shown in Table 3, we can see that the transfer learning methods such as DTL, DFF and ResNet50-baseline achieve worse performance than

Table 3 Accuracies (%) of models with single source domain

Source domain	Catech256	Amazon	Webcam	Dslr	Breast	Tucc
DTL	82.99	77.37	73.63	74.72	80.34	81.43
Co-Tuning	**83.93**	**80.88**	79.09	78.93	82.18	86.27
DFF	75.32	76.96	76.99	76.46	73.13	76.02
SLA	79.40	80.49	79.87	79.25	80.03	79.71
AM-DDTL	80.49	80.40	**80.19**	79.40	81.12	83.46
Resnet50	74.70	74.87	76.04	77.48	80.05	85.58
DHFF	78.53	76.14	79.54	**82.50**	**84.14**	**87.20**

Table 4 Accuracies (%) of models with breast ultrasound auxiliary source domain

Source domain	Catech256	Amazon	Webcam	Dslr
Auxiliary Source Domain	Breast ultrasound			
DFF	76.70	77.20	76.17	77.17
Resnet50	77.98	76.52	77.00	82.93
SLA	81.74	82.21	82.21	80.03
AM-DDTL	**82.83**	82.52	81.27	81.43
DHFF	82.68	**83.44**	**84.39**	**86.66**

Table 5 Accuracies (%) of models with thyroid ultrasound auxiliary source domain

Source domain	Catech256	Amazon	Webcam	Dslr	Breast
Auxiliary Source Domain	Thyroid Ultrasound Cine-clip				
DFF	76.36	75.78	76.59	78.75	74.76
Resnet50	83.06	79.19	75.23	85.50	87.33
SLA	82.83	82.05	82.68	81.12	80.49
AM-DDTL	83.61	82.52	81.59	82.05	83.93
DHFF	**84.39**	**82.74**	**84.73**	**86.73**	**88.92**

SLA, AM-DDTL and DHFF by using the Office 31 (distant domains) as the source domain, because the source domain and the target domain have huge distribution gap, which leads to negative transfer. SLA performs well because it selects knowledge valuable for distant transfer learning based on instance and AM-DDTL benefits from the CBAM (Convolutional Block Attention Module) attention mechanism which making use of the channel and the spatial attention for better feature extraction. In addition, our proposed DHFF model performs well on different datasets, DHFF can not only extract appropriate high-level semantic features, but also use content loss to avoid negative transfer caused by excessive feature fusion between two different domains. It achieves the highest accuracy (87.20%) using the Thyroid Ultrasound Cine-clip dataset (TUCC) as the source domain, which has the highest correlation with the target domain. Moreover, compared with DTL, Co-Tuning, applying the collaboratively supervise the fine-tuning process, can ameliorate the negative transfer effect of traditional fine-tuning on distant domains, however the method is still not enough to obtain more valuable

transfer knowledge on the source domain (Breast and Tucc) which has high correlation and the same categories with the target domain. And when the target domain dataset is small, the Co-Tuning is easy to overfit to the limited labeled train data. Furthermore, compared with the DHFF, DTL and ResNet50-baseline, the feature extraction structure of DFF, SLA is so simple and cannot extract enough high-level semantic information in ultrasound imaging dataset, so its results are not satisfactory. Although AM-DDTL alleviates this problem, its feature extraction relying on attention mechanism is still insufficient. It is proved that high-level semantic information plays an important role in distant domain transfer learning.

When the correlation between the source domain and the target domain is low, the transferred knowledge is not helpful for the task of the target domain, and may even cause negative transfer. In experiment 2, we adopt breast ultrasound images as the auxiliary domain and natural images as the source domain. The results in Table 4 shows that the auxiliary domain which closer to the target domain may help to narrow the gap between the source and target domain and is beneficial to improve the performance of the target task. In this experiment, compared with ResNet50-baseline, the SLA, DFF and AM-DDTL which belongs to DDTL methods can alleviate distant domain negative transfer. But they cannot sufficiently extract high-level semantic features and the performance is lower than that of the method proposed in this paper. In contrast, DHFF achieves the highest accuracy (86.66%) when breast ultrasound images are served as auxiliary domain. It further proves DHFF can effectively transfer valuable knowledge compared with other models.

In addition, we also experiment with thyroid ultrasound images as the auxiliary domain in experiment 3. The results shown in Table 5 prove that using the thyroid ultrasound images as the auxiliary domain which share the most useful knowledge with the target domain can maximize the model classification accuracy. This shows that the auxiliary domain which is more relevant to the target domain can help distant transfer learning. Surprisingly, when the breast dataset and thyroid datasets are adopted as source domain and auxiliary domain respectively, the highest classification accuracy (88.92%) is achieved. This shows that when the thyroid dataset is used as a bridge, the breast dataset with relatively strong correlation can provide the most valuable knowledge to the target domain. On the contrary, SLA is not satisfactory in the high correlation source domains, which proves that the instances-based method may ignore the potentially valuable transfer knowledge. Moreover, although highly correlation auxiliary domain facilitates the valuable information transfer, DFF and AM-DDTL which extract low level features to fusion cannot provide effective information for the classifier.

Finally, we conduct experiments with a small thyroid ultrasound dataset (716 images). The results are shown in Table 6. Our method DDHF can still achieve best results (86.27%), which benefits from the transfer of valuable knowledge from the source and auxiliary domains to the target domain. However, negative transfer occurs on all five source domains based on the DFF, SLA and AM-DDTL model. The reason is that due to the simple feature extractor and few samples, DFF and SLA cannot effectively learn high-level semantic features from different domains, making the features of different domains excessive overlap, ultimately lead to the model underfitting the target domain data. In addition, the attention mechanism model AM-DDTL which relies on a large amount of data cannot achieve good results.

Further, we analyzed the samples which misclassified by the model. Fig. 7 shows that benign samples are misclassified as malignant, and Fig. 8 shows that malignant samples are misclassified as benign.

In Fig. 7, each sample shows some malignant features, such as no cyst, hypoechoic, rough edge. They are classified as Ti-rads4 by radiologist. In Fig. 8(a), the internal echo attenuation

Table 6 Accuracies (%) of models with few samples

Source domain	Catech256	Amazon	Webcam	Dslr	Breast
Auxiliary Source Domain	Thyroid Ultrasound Cine-clip				
DFF	70.24	69.26	71.30	69.22	70.33
Resnet50	80.75	79.46	83.63	85.05	85.11
SLA	76.47	75.49	75.49	78.43	77.45
AM-DDTL	77.56	77.07	78.53	79.02	81.46
DHFF	85.23	80.97	85.09	85.27	**86.27**

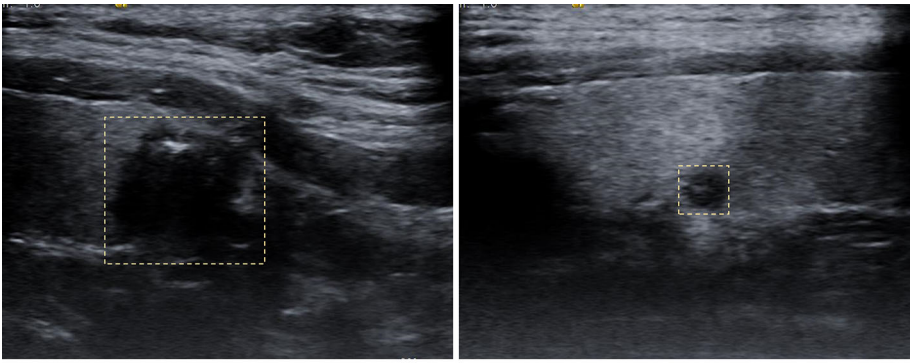


Fig. 7 benign samples misclassified as malignant

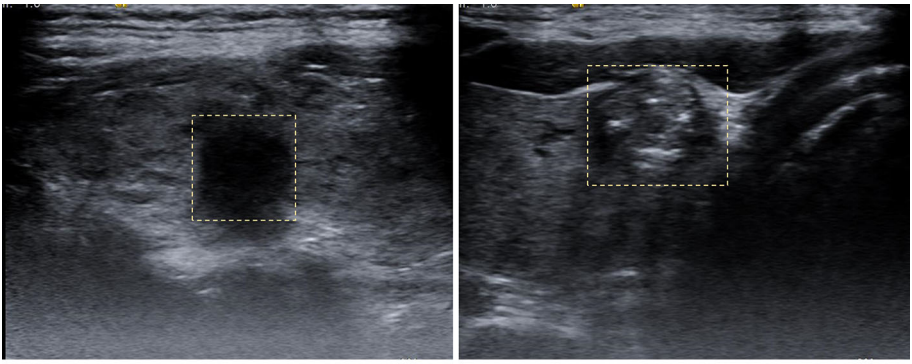


Fig. 8 malignant samples misclassified as benign

is too great, which is close to cyst. In Fig. 8(b), there is some wall structure, which is benign feature.

From the analysis, it can be seen that overlap in the image characteristics between benign and malignant thyroid nodule is one of the causes of classification errors. The limited number of training samples also affects the generalization performance of the model. In addition, the static image only reflects one section of the nodule, and cannot reflect the whole picture of the

nodule. Using the dynamic image or multiple static images of a nodule can help to improve the accuracy.

6 Conclusions

This paper proposes a novel method to solve the problem of distant domain transfer learning. The purpose is to transfer valuable information in the distant domain to improve the task performance in target domain. A lightweight feature extractor is proposed to extract high-level semantic features in both source domain and target domain. The MMD loss is to narrow the distribution distance of source domain and target domain. Then, an encoder-decoder network and content loss are proposed to maintain the diversity of the source domain and the target domain. Experiments show that this method can effectively transfer knowledge from distant domain and improve the accuracy of thyroid ultrasound image classification.

When the samples in the target domain are difficult to collect and annotate, our method can effectively transfer the knowledge in the distant domain, which is conducive to improving the task in the target domain. This makes the deep learning model more applicable to small sample scenarios. Generally, we think that training data and test data are independent and identically distributed (IID), but this is not the case in practical applications. For example, the thyroid ultrasound images collected by different machines usually are not IID, which brings difficulties to the learning task. The proposed method can partly solve the problem by transfer the knowledge of the distant domain, to improve the generalization of the model. For example, in the experiment, using breast ultrasound image as the source domain can improve the classification performance of thyroid ultrasound image.

Due to the high cost of collecting and labeling data in medical imaging, algorithms in this field usually use small data for training. To improve performance, transfer the model trained in big data to the target task is a simple and effective method. But if the source and target domain are not similar enough, it will be difficult to use transfer learning directly. From the perspective of target task, we hope to shorten the distance between source domain and target domain; from the perspective of generalization, we hope to retain the differences between the source domain and the target domain, and transfer high-level semantic features which is helpful to the target task. This is a trade-off. The method proposed in this paper makes a useful exploration to solve this problem. Although some good results have been achieved, there are still several limitations: 1) Lack of means to guide the selection of the source domain to maximize the performance of the target domain. 2) The auxiliary domain needs to use unlabeled ultrasound images with high similarity to the target domain, which is relatively difficult to collect.

In future research, we will further analyze which knowledge transferred by the model is helpful to improve performance of target domain tasks and which knowledge will have an adverse impact on target tasks. So as to further improve the performance of the model, reveal the internal working principle of the model and improve the interpretability of the model.

Acknowledgements This work is supported, in part, by Shandong Natural Science Foundation of China; the Grant numbers is ZR2020MH290.

Declarations

Conflict of Interest The authors declared that they have no conflicts of interest to this work.

References

1. Huang Q, Zhang F, Li X (2018) Machine learning in ultrasound computer-aided diagnostic systems: A survey. BioMed research international 2018
2. Singh N, Jindal A (2012) Ultra sonogram images for thyroid segmentation and texture classification in diagnosis of malignant (cancerous) or benign (non-cancerous) nodules. Int. J. Eng. Innov. Technol 1:202–206
3. Bibicu D, Moraru L, Biswas A (2013) Thyroid nodule recognition based on feature selection and pixel classification methods. J of digital imaging 26(1):119–128
4. Tsantis S, Dimitropoulos N, Cavouras D, Nikiforidis G (2009) Morphological and wavelet features towards sonographic thyroid nodules evaluation. Comput Medical Imaging and Graphics 33(2):91–99
5. Chang CY, Chen SJ, Tsai MF (2010) Application of support-vector-machine-based method for feature selection and classification of thyroid nodules in ultrasound images. Pattern recognit 43(10):3494–3506
6. Song W, Li S, Liu J, Qin H, Zhang B, Zhang S, Hao A (2018) Multitask cascade convolution neural networks for automatic thyroid nodule detection and recognition. IEEE j of biomed and health inform 23(3):1215–1224
7. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. nat 521(7553):436–444
8. Pan SJ, Yang Q (2009) A survey on transfer learning. IEEE Trans on knowledge and data eng 22(10):1345–1359
9. Kora P, Ooi CP, Faust O, Raghavendra U, Gudigar A, Chan WY, Acharya UR (2021) Transfer learning techniques for medical image analysis: A review. Biocybernetics and Biomedical Engineering
10. Tan B, Song Y, Zhong E, Yang Q (2015 August) Transitive transfer learning. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining pp 1155-1164
11. Tan B, Zhang Y, Pan S, Yang Q (2017 February) Distant domain transfer learning. In: Proceedings of the AAAI conference on artificial intelligence 31(1)
12. Niu S, Liu M, Liu Y, Wang J, Song H (2021) Distant domain transfer learning for medical imaging. IEEE J of Biomed and Health Inform 25(10):3784–3793
13. Qin S, Guo X, Sun J, Qiao S, Zhang L, Yao J, Zhang Y (2021) Landslide Detection from Open Satellite Imagery Using Distant Domain Transfer Learning. Remote Sensing 13(17):3383
14. Qin P, Wu K, Hu Y, Zeng J, Chai X (2019) Diagnosis of benign and malignant thyroid nodules using combined conventional ultrasound and ultrasound elasticity imaging. IEEE J of Biomed and Health Inform 24(4):1028–1036
15. Wang Y, Yue W, Li X, Liu S, Guo L, Xu H, Yang G (2020) Comparison study of radiomics and deep learning-based methods for thyroid nodules classification using ultrasound images. Ieee Access 8:52010–52017
16. Shi G, Wang J, Qiang Y, Yang X, Zhao J, Hao R, Kazihise NGF (2020) Knowledge-guided synthetic medical image adversarial augmentation for ultrasonography thyroid nodule classification. Comput Methods and Programs in Biomed 196:105611
17. Ma J, Wu F, Jiang TA, Zhao Q, Kong D (2017) Ultrasound image-based thyroid nodule automatic segmentation using convolutional neural networks. Int j of comput assisted radiology and surgery 12(11):1895–1910
18. Zhou H, Wang K, Tian J (2020) Online transfer learning for differential diagnosis of benign and malignant thyroid nodules with ultrasound images. IEEE Trans on Biomed Eng 67(10):2773–2780
19. Chi J, Walia E, Babyn P, Wang J, Groot G, Eramian M (2017) Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. J of digital imaging 30(4):477–486
20. Raghu M, Zhang C, Kleinberg J, Bengio S (2019) Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems 32
21. Xie M, Jean N, Burke M, Lobell D, Ermon S (2016 March) Transfer learning from deep features for remote sensing and poverty mapping. In: Thirtieth AAAI Conference on Artificial Intelligence
22. Borgwardt KM, Gretton A, Rasch MJ, Kriegel HP, Schölkopf B, Smola AJ (2006) Integrating structured biological data by kernel maximum mean discrepancy. Bioinform 22(14):e49–e57
23. 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
24. Turchenko V, Chalmers E, Luczak A (2017) A deep convolutional auto-encoder with pooling-unpooling layers in caffe. arXiv preprint arXiv:1701.04949
25. Griffin G, Holub A, Perona P (2007) Caltech-256 object category dataset
26. Zhao Y, Ali H, Vidal R (2017) Stretching domain adaptation: How far is too far?. arXiv preprint arXiv:1712.02286
27. Thyroid Ultrasound Cine-clip. https://stanfordaimi.azurewebsites.net/datasets/a72f2b02-7b53-4c5d-963c-d7253220bfd5

28. Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks?. Advances in neural information processing systems 27
29. You K, Kou Z, Long M, Wang J (2020) Co-tuning for transfer learning. Adv in Neural Inform Process Syst 33:17236–17246
30. Deng J (2009) A large-scale hierarchical image database. Proc. of IEEE Computer Vision and Pattern Recognition 2009