Progressive Transfer Learning and Adversarial Domain Adaptation for Cross-Domain Skin Disease Classification

Yanyang Gu, Zongyuan Ge, *Member, IEEE*, C. Paul Bonnington, *Senior Member, IEEE*, and Jun Zhou, *Senior Member, IEEE*

***Abstract*—Deep learning has been used to analyze and diagnose various skin diseases through medical imaging. However, recent researches show that a well-trained deep learning model may not generalize well to data from different cohorts due to domain shift. Simple data fusion techniques such as combining disease samples from different data sources are not effective to solve this problem. In this paper, we present two methods for a novel task of cross-domain skin disease recognition. Starting from a fully supervised deep convolutional neural network classifier pre-trained on ImageNet, we explore a two-step progressive transfer learning technique by fine-tuning the network on two skin disease datasets. We then propose to adopt adversarial learning as a domain adaptation technique to perform invariant attribute translation from source to target domain in order to improve the recognition performance. In order to evaluate these two methods, we analyze generalization capability of the trained model on melanoma detection, cancer detection, and cross-modality learning tasks on two skin image datasets collected from different clinical settings and cohorts with different disease distributions. The experiments prove the effectiveness of our method in solving the domain shift problem.**

***Index Terms*—Automatic melanoma detection, dermoscopy image, cycle-GAN, deep learning, transfer learning, domain adaptation.**

I. Introduction

SKIN cancer is one of the most frequently diagnosed human malignancy, especially among light pigmented skin [1]. Malignant melanoma is the most deadly category of all skin cancers, although it is not the most prevalent type. The number of melanoma incidences (287,723) is only about a quarter of non-melanoma (1,042,056) from 185 countries in 2018, but the melanoma mortality (60,712) is almost the same as all other non-melanoma categories (65,155) combined [2]. Fortunately, early screening of melanoma can greatly increase the chance of cure.

In both on-site and teledermatology diagnosis, expertise in skin cancer recognition is acquired by a mixture of knowledge-based education and exposure to lesions to learn their visual characteristics. Due to the subjectivity and level of training received by different clinicians, some early stage skin cancer may not be easily recognizable. Automated image based diagnosis systems are developed to address this issue and reduce the workload of clinicians, as a subsidiary method to skin cancer detection.

Such automated diagnostic systems can be divided into two categories: early classical machine learning methods and recent deep neural networks (DNN) based methods. Early systems were trained on relatively small scale datasets [3]. They usually followed the steps of pre-processing [4], segmentation [5], feature extraction [6], classification [7] and sometimes post-processing [8]. Each of the steps requires to be manually designed with expertise, especially the extraction of hand-crafted features. It is normally hard to apply such classical machine learning methods to new scene. Recently, deep neural networks were used to analyze pathology and clinical images and showed great success [9]–[11], including for skin disease recognition [7], [12]–[17], which alleviates the drawbacks of classical machine learning methods. Most of these methods [7], [12]–[15], [18] built skin disease recognition models in single modality [19], [20]. Ge *et al.* [16] considered multi-modality circumstances of skin imaging, which included clinical photography of a macro view and dermoscopy imaging of a micro view of the mole. Results show that multi-modality leads to better diagnosis results by providing complementary visual features.

Though skin disease recognition based on deep learning is promising [21], models trained on one skin dataset may not work equally well on different skin disease datasets. As shown in Table I, the classification accuracy of 7 overlapping skin diseases

TABLE I
Results of CNN Models on MoleMap and HAM Datasets. This Table Also Illustrates the Motivation of This Paper

Dataset Name	Train%	Test-HAM%	Test-MoleMap%
HAM	0.952	0.907	0.310
MoleMap	0.900	0.535	0.795
MoleMap+HAM	0.913	0.902	0.792

Manuscript received March 12, 2019; revised July 4, 2019 and August 16, 2019; accepted September 11, 2019. Date of publication September 23, 2019; date of current version May 6, 2020. *(Yanyang Gu and Zongyuan Ge contributed equally to this work.) (Corresponding authors: Jun Zhou; Zongyuan Ge.)*

Y. Gu and J. Zhou are with the Griffith University, Nathan 4111, Australia (e-mail: yanyang.gu@griffithuni.edu.au; jun.zhou@griffith.edu.au).

Z. Ge is with the Monash University, Clayton 3800, Australia (e-mail: zongyuan.ge@monash.edu.au).

C. P. Bonnington is with the Monash University, Clayton 3800, Australia, and also with the Airdoc, Shanghai 200000, China (e-mail: paul.bonnington@monash.edu).

Digital Object Identifier 10.1109/JBHI.2019.2942429

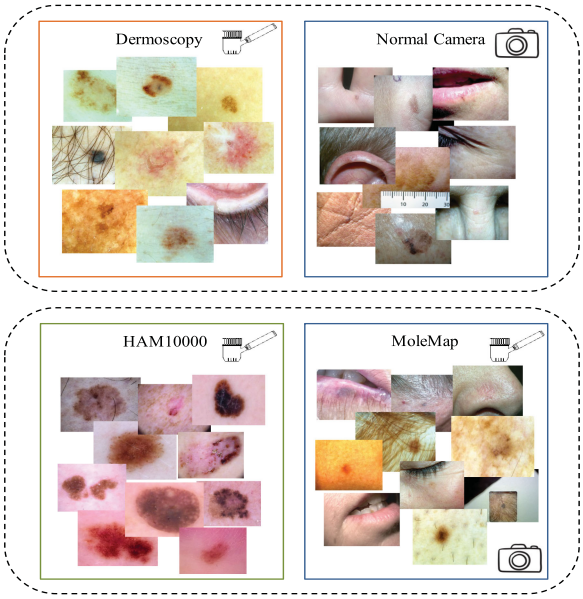


Fig. 1. Image samples of two different modalities in HAM and MoleMap datasets. HAM10000 samples are all dermoscopic images while MoleMap samples contain both dermoscopic and normal camera images. There clear exists domain shift between these two datasets due to different capturing settings.

between two datasets using pre-trained ResNet-152 models [22] were first fine-tuned on one dataset, and tested on another. Here two skin datasets MoleMap and HAM (check Section III for more details) are being used. Column 2 to 4 show the training and testing accuracy across various datasets. In all comparisons, we observe that the testing accuracy of a model trained on a different dataset is significantly lower than that on the original dataset. Even training the model using the combination of both datasets, it still underperforms the results on the original dataset. This is due to the factors that different datasets are collected from a variety of clinical settings and patient cohorts. As shown in the second row of Fig. 1, samples of two different datasets are captured using different imaging devices (e.g. dermoscope for HAM while both dermocope and normal camera for MoleMap) and different labelling standards (e.g. how to annotate possible melanoma and definite melanoma). These issues result in domain shift and change of sample distributions, which in turn, cause inferior performance.

The primary aim of this paper is to show how a deep learning model trained on one dataset can be generalized to another dataset obtained from a different resource. In real world medical applications, the generalization capability is very important in introducing less bias into the diagnosis system. We propose two methods to enable a model to generalize more effectively to new samples, and alleviate domain shift between the source and target datasets. The first method is parameter based progressive transfer learning [23] which learns new information in a model-adaptation manner. The second method is adversarial based domain adaptation to map the attributes of the skin imagery from the source domain to the target domain, which is a pixel-wise image synthesizing adaptation instead of parameter-based model adaptation. Unlike the previous transfer learning method, the adversarial based domain adaptation method is semi-supervised and does not require any labels from the new dataset for the adaptation process.

The contributions of this work are:

1) We propose to adopt a two-step progressive transfer learning method for connecting task-specific skin cancer classification. We treat one skin dataset as the intermediate set and the other one as the final target set. We first pre-train a Convolutional Neural Network (CNN) model on ImageNet [24], and fine-tune it on the intermediate set. Then we fine-tune the model on the target testing skin dataset so as to take advantage of the pre-learned knowledge step by step. The experiments show that progressive transfer learning achieves better performance than single step transfer learning.
2) We propose to adopt cycle-consistent adversarial learning [25] as a domain adaptation technique to perform skin imaging attributes translation (such as illumination, light consistency around the lesion boundary etc.) from the source to the target dataset. The transformed images are beneficial to generalize the trained model across other datasets. Comparing with the model trained on original datasets only, experimental results demonstrate that the model trained with domain adaptation transformed data is capable of achieving better disease recognition.[1]
3) We run cross-sectional experiments using multiple model training cohorts to evaluate the generalization capability of trained model to external datasets. We explore both general imaging and modality-specific transformation using a cycle-consistent adversarial domain adaptation technique. Empirical results show that this technique is capable of transforming images between different modalities, such as from dermoscopy images to normal camera images and vice versa, so as to improve the classification performance.

The rest of the paper is organized as follows. We present the related works in Section II. The details of our method are introduced in Section IV. Section V describes experimental evaluations and comparison with ground truth methods to verify the effectiveness of our method. Finally, some insights and conclusions from the experiments are discussed in Section VI.

II. Related Work

It is common to see a drop of performance when CNN models are trained and tested on different datasets for the same task, as shown in Table I. The main reason is that different datasets (domains) have different data distributions, which is typically known as domain shift [26]. This issue can be addressed by domain adaptation, which either finds a common domain or diminishes the distribution change and domain shift between different domains [27]. Alternatively, solutions based on transfer learning or knowledge transfer aim at transferring the model, which contains knowledge learned from one dataset, to adapt

[1] melanoma-v.s.-non-melanoma and cancer v.s.non-cancer classification task.

to the new domain [23]. These two methods have attracted substantial attention in the machine learning and computer vision communities, especially with the recent booming of deep learning.

Domain adaptation and transfer learning methods adopt three different designs of loss functions, including classification loss, adversarial loss and discrepancy loss [27]. The first design, classification loss, is adopted by transfer learning to optimize the deep networks. Methods using the second design, adversarial loss, can be divided into generative models [28] and non-generative models. The former synthesizes samples by means of data-driven techniques, while the latter learns the mappings from the source domain to the target domain through a domain-confusion loss. The third design, namely discrepancy loss, assumes that reconstruction of the source or target data based on discrepancy loss can facilitate domain adaptation. This can be achieved by stacking autoencoders [29] or generative adversarial reconstruction [25], [30], [31]. The difference between adversarial loss and discrepancy loss based methods is that the former method aims to diminish the shift between two domains while the latter method tries to encourage a common feature space [27].

In order to be accordant with our proposed scheme, the related methods are reviewed in terms of a narrow scope of transfer learning and adversarial domain adaptation.

A. Transfer Learning

Transfer learning is a broad concept [23]. A narrow scope of transfer learning was exploited in the early stage of deep learning, in which unsupervised pre-train of representations were transferred to facilitate prediction [32], [33]. In various deep network architectures, features in the shallow-layers are usually general while high-level features are specific. Yosinski *et al.* [34] proved that initializing a network with parameters of a model transferred even from distant tasks improves the performance compared with random parameters. Therefore, in many deep learning applications without sufficient training data [35], [36], the shallow layers of the pre-trained model are kept to retain the generalized transferable knowledge. The deep layers are fine-tuned via back-propagation to obtain data specific semantically meaningful features [26].

B. Adversarial Domain Adaptation

Based on the generative adversarial networks (GANs), Isola *et al.* [30] introduced conditional adversarial networks into image-to-image translation, regularizing the mapping between inputs and outputs. However, this requires paired training samples to train a model, which means fake paired images have to be generated as training samples. This is not only a burden for computation but is not reliable either. Therefore, GANs related unpaired image-to-image translation method (cycle-GAN) is proposed. It adds an inverse mapping to enforce the generated outputs to be converted back to the input space, thereby making a cycle loss consistency process [25]. Cycle-GAN is based on the assumption that two domains are highly dependent. The method collapses when translating different domains with large structure changes, such as the reconstruction of a video of Barack Obama from Donald Trump [31]. To address this issue, temporal consistency and spatial cycle consistency are considered together to translate videos from one domain to a target domain yet still preserving the input domain style (Recycle-GAN) [31]. This in turn is an advantage of our method since it needs to retain structural information and produce as little structure loss as possible for accurate melanoma detection.

III. Datasets

We choose two skin disease datasets MoleMap and HAM to verify our proposition and proposed methods. MoleMap dataset represents the data collected from a clinical environment and followed tele-dermatology labelling standard. HAM dataset is the largest publicly available skin dataset collected from ISIC archive,[2] and it is most technically advanced and accessible resource for digital dermatoscopy. Those two datasets consist of samples from different disease and corhort distributions, which lead to ideal experiments to study domain shift for dermatology. The details about disease distribution of these two datasets are listed in Tables II and III. The first column indicates the name of the disease, the second shows the number of samples for each disease, the third row shows the percentage of the disease, and the last row indicates the ratio between the macro images (clinical images) and the micro images (dermoscopy images).

MoleMap:[3] This dataset has 102 451 images with 25 skin conditions, including 22 benign categories and 3 cancerous categories. The cancerous categories include melanoma (pink melanoma, normal melanoma and lentigo melanoma), basal cell carcinoma and squamous cell carcinoma. Each lesion has two images: a close-up image taken at a distance of 10 cm from the lesion (called the macro) and a dermoscopic image of the lesion (called the micro). Images were chosen for the study according to four criteria. First, each image is associated with a disease specific diagnosis (e.g. blue nevus). Second, there are at least 100 images with the same diagnosis. Third, the image quality is acceptable (e.g with good contrast). Last, the lesion occupied most of the image without a lot of surrounding tissue.

HAM10000 (HAM): This dataset consists of 10015 dermatoscopic images in 7 categories, including 5 benign categories and 2 cancerous categories (i.e. melanoma and basal cell carcinoma). The images were collected over a period of 20 years from Australia and Austria. The Australian data are all digital, but the data from Austria include both digital dermatoscopic images and non-digital diapositives, where the latter type was digitized by scanning and manual correction.

Besides the differences of acquiring process and clinical settings between these two datasets, HAM contains only dermoscopic images with 50% of lesions being confirmed by pathology, while MoleMap includes both clinical and dermoscopy images with tele-dermatology verification but no pathological confirmation.

[2]https://isic-archive.com/
[3]http://molemap.co.nz.

TABLE II
MOLEMAP (25 CLASSES) DATASET DETAILS

Id	Disease names	Number	% of total	ma/mi
0	Lentigo	386	0.40%	0.763
1	Dysplastic naevus	9033	9.25%	0.86
2	Compound naevus	5103	5.23%	0.711
3	Regressing naevus	176	0.18%	0.778
4	Irritated (Dermal) naevus	113	0.12%	0.948
5	Acral naevus	193	0.20%	0.949
6	Blue naevus	520	0.53%	0.837
7	Dermal naevus	938	0.96%	0.872
8	Junctional naevus	1758	1.80%	0.703
9	Keratotic lesion & Warty keratosis	2020	2.07%	0.846
10	Lichenoid keratosis	208	0.21%	0.748
11	Porokeratosis	204	0.21%	0.789
12	Actinic cheilitis	217	0.22%	0.855
13	Vascular & Angioma & Hemangioma & Cheery angioma & Venous Lake	855	0.88%	0.835
14	Dermatofibroma	2310	2.37%	0.881
15	Scar	382	0.39%	0.777
16	Sebaceous hyperplasia	606	0.62%	0.825
17	Melanoma	6608	6.77%	0.779
18	Lentigo Melanoma & Sus	293	0.30%	0.754
19	Basal Cell Carcinoma	16266	16.66%	0.763
20	Squamous Cell Carcinoma	2261	2.32%	0.735
21	Bowens disease	1819	1.86%	0.783
22	(Bowenoid) Actinic Keratosis	28051	28.73%	0.879
23	Solar lentigo	3045	3.12%	0.704
24	Seborrheic keratosis	14280	14.62%	0.821

TABLE III
HAM10000 DATASET DETAILS

Disease names	Number	% of total	mi/ma
Actinic keratosis / Bowens disease	327	3.27%	1
Basal cell carcinoma	514	5.13%	1
Benign keratosis	1099	10.97%	1
Dermatofibroma	115	1.15%	1
Melanoma	1113	11.11%	1
Melanocytic nevus	6705	66.95%	1
Vascular lesion	142	1.42%	1

IV. METHODS

As shown in Table I, deep CNN models trained on the dataset obtained from one clinical setting or on directly combined datasets can not work effectively on datasets obtained from other settings due to distribution shift of the data (domain shift). In order to tackle this problem, we investigate two options, transfer learning and domain adaptation, to make data obtained from different settings equally useful to train a robust model.

In this section, we introduce a progressive transfer learning structure as the first solution. This is followed by a cycle consistent adversarial networks (cycle-GAN) for image synthesizing adaptation. Then we briefly explain dataset domain adaptation and modality domain adaptation using the synthesized data, as the second solution to the distribution drift issue. Finally, discussion on binary and multi-class classification is given.

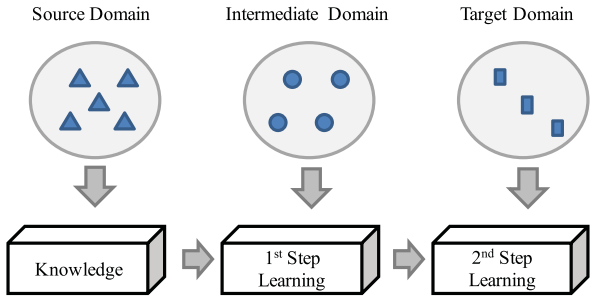


Fig. 2. Illustration of the progressive transfer learning process. The transferable knowledge is transferred to the intermediate domain before to the target domain.

A. Progressive Transfer Learning

In image representation and label space, it is common to observe a large data distribution shift [37], [38] from large-scale source domain datasets such as ImageNet [24] to medium/small size target domain datasets such as Pascal VOC [39] and CIFAR100 [40]. This may lead to inferior generalization ability when a model is trained only in the source domain. Transfer learning [23] is a technique to circumvent this issue. In the deep learning setting, the most widely used approach for transfer learning is to train a model in the source domain $\mathcal{S}$ (where the categories are often diverse and abundant) and then adapt it to the target domain $\mathcal{T}$. This method has demonstrated great success in various deep CNN models such as VGGNet [41], ResNet [22], DenseNet [42].

We first define notations and terminology in transfer learning. Let $\mathcal{T}$ and $\mathcal{S}$ be the target and the source domains, respectively. Let $\mathcal{X}_S$ be the sample space of the source domain and $\mathcal{X}_T$ be the sample space of the target domain, respectively. $\mathcal{P}_S := \mathcal{P}_{S,\theta_S}$ with parameters θ_S is the model whose empirical risk on the source dataset shall be minimized during training. Given these definitions, parameter based transfer learning can be defined as follows:

$$\mathcal{H} = \{h_{\mathcal{ST}} : \mathcal{P}_\mathcal{S} \rightarrow \mathcal{P}_\mathcal{T}\} \tag{1}$$

where $\mathcal{H}$ is the hypothesis to perform parameter transfer learning from $\mathcal{S}$ to $\mathcal{T}$, that is, to fine-tune the pre-trained model to adapt to the target domain testing dataset. An assumption here is that $\mathcal{P}_\mathcal{S}$ of the parameter space is effective in the source region, and a part of the parameter space is shared with the target domain.

Parameter-based transfer learning can be extended to multiple-step progressive transfer learning when more than two datasets are involved in the adaptation process:

$$h_{\mathcal{SMT}} : \mathcal{P}_\mathcal{S} \rightarrow \mathcal{P}_\mathcal{M} \rightarrow \mathcal{P}_\mathcal{T} \tag{2}$$

where $\mathcal{P}_\mathcal{M}$ is the intermediate domain as shown in Fig. 2. In this paper, we assume that the intermediate domain $\mathcal{X}_M$ and the target domain $\mathcal{X}_T$ belong to the same category as the

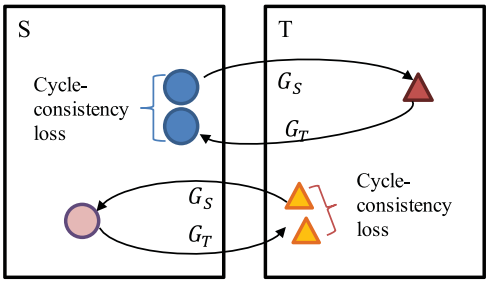


Fig. 3. Illustration of the cycle-GAN model. Cycle consistency loss is introduced so as to allow the generated sample converted back to its original domain.

target domain (they are both skin disease dataset of the same categories) Furthermore, these datasets have to abide to the following relationship:

$$d(\mathcal{X}_M, \mathcal{X}_T) < d(\mathcal{X}_S, \mathcal{X}_T) \tag{3}$$

where d denotes the difference between two domains.

B. Generative Adversarial Networks

As an alternative to parameter based transfer learning, synthetic images can be used to facilitate the domain adaptation task between multiple datasets. In this work, we propose to address the multi-domain adaptation problem using Generative Adversarial Network (GAN) based methods [25]. We first review GAN and cycleGAN models.

GAN: GAN has generated impressive results in a wide range of image generation and translation tasks [27]. Isola *et al.* [30] applied conditional GAN to learn pairwise image-to-image translation. The key idea of GAN is generating synthetic images that are similar to real images [28]. An adversarial loss is defined to encourage the generated images to be distinguishable from real ones. A discriminator D is trained adversarially to ensure the generated samples from the generator $G(x_i)$ $(x_i \in \mathcal{X}_S)$ to be different from the target $y_i \in \mathcal{X}_T$, thereby maximizing the discrimination while minimizing the mapping error as defined in the following loss function:

$$\begin{aligned}\mathcal{L}_{GAN}(G, D) = &\mathop{\mathbb{E}}_{\mathbf{y}}[logD(y)] \\ &+ \mathop{\mathbb{E}}_{\mathbf{x}}[log(1 - D(G(x))].\end{aligned} \tag{4}$$

Apart from the adversarial loss, traditional construction loss is found to be beneficial to be mixed with GAN loss:

$$\mathcal{L}_o(G) = \sum_i ||y_i - G(x_i)||^2. \tag{5}$$

CycleGAN [25]: Since pairwise image data is difficult to be acquired, cyleGAN [25] was proposed to learn mapping functions between two unpaired image domains $\mathcal{S}$ and $\mathcal{T}$ as shown in Fig. 3, where the one way arrow can be indicated as the normal GAN and the arrow pairs denote the cycleGAN. In addition to the generative adversarial loss, a cycle-consistent loss is introduced. Generative adversarial loss guarantees the synthesized samples generated in the target domain (on the right of Fig. 3) to be different from their sources (on the left of Fig. 3), while cycle-consistent loss forces the synthesized images to be similar if they are converted back to the original source domain. CycleGAN works well when elements in both domains are highly dependent, although otherwise it fails as illustrated in Fig. 6. CycleGAN firstly applies generators to both mapping directions, i.e. from the source domain to the target domain $G_{ST}(x)$ and reversely $G_{TS}(y)$. In addition, it introduces a cycle loss [43] to keep cycle consistency, so that for any image sample $x \in \mathcal{X}_S$, the generated sample $G_{ST}(x)$ that is mapped to the target domain $\mathcal{X}_T$ can be converted back to the source domain, i.e. $G_{TS}(G_{ST}(x)) \approx x$. Similarly, for any image sample $y \in \mathcal{X}_T$, $G_{ST}(G_{TS}(y)) \approx y$, as illustrated in Fig. 3. The objective of cycle loss starting from $\mathcal{D}^S$ is:

$$\mathcal{L}^S_{cycle}(G_{ST}, G_{TS}) = \mathop{\mathbb{E}}_{\mathcal{X}_S} ||x - G_{TS}(G_{ST}(x))||_1, \tag{6}$$

where G_{ST} denotes generator from source domain $\mathcal{X}_S$ to target domain $\mathcal{X}_T$, while G_{TS} denotes generator from target domain $\mathcal{X}_T$ to source domain $\mathcal{X}_S$. Therefore, the full objective function of cycleGAN is:

$$\begin{aligned}&\mathcal{L}_{cycleGAN}(G_{ST}, G_{TS}, D_S, D_T) \\ &= \mathcal{L}^S_{GAN}(G_{ST}, D_S) + \mathcal{L}^T_{GAN}(G_{TS}, D_T) \\ &\quad + \lambda\mathcal{L}^S_{cycle}(G_{ST}, G_{TS}) + \lambda\mathcal{L}^T_{cycle}(G_{TS}, G_{ST}),\end{aligned} \tag{7}$$

where λ is a hyper-parameter that balances the influence of GANs loss and cycle consistency loss, D_S and D_T denote the discriminators applied to source domain and target domain, respectively.

C. Synthesizing Domain Adaptation

Although deep CNNs are promising for disease diagnosis from medical images [13], [14], it is pointed out that CNN models trained on a dataset acquired from one hospital does not work equally well in datasets collected from different hospitals. Parameter based transfer learning introduced in Section IV-A shows how to perform cross-data adaptation in parameter space. In this section, we propose to use cycleGAN based method to benefit cross-data domain and modality adaptation. Fig. 4 illustrates both processes of Data Domain Adaptation and Modality Domain Adaptation, where the notations on the upper left corner of each block have the following meaning: **M/D** denotes all images or dermoscopy images in MoleMap, $\mathbf{M}_T$/$\mathbf{D}_T$ denotes the synthesized images after cycleGAN, $\mathbf{T}_o$ denotes the original target dataset, and $\mathbf{T}_A$ denotes the augmented dataset. To start, m samples are randomly selected from the source domain and the target domain to train the cycleGAN model, with which a small part of the source domain images are synthesized into the target domain. The number of the synthesized target images is related to the size of the original target dataset n, which is $p\%n$. Finally, these $p\%n$ synthesized images are combined with the original n target samples to form the training set for the classification model.

1) Dataset Domain Adaptation: HAM and MoleMap datasets represent different patient cohorts, thus there is internal distribution shift between them. To make a model generalized

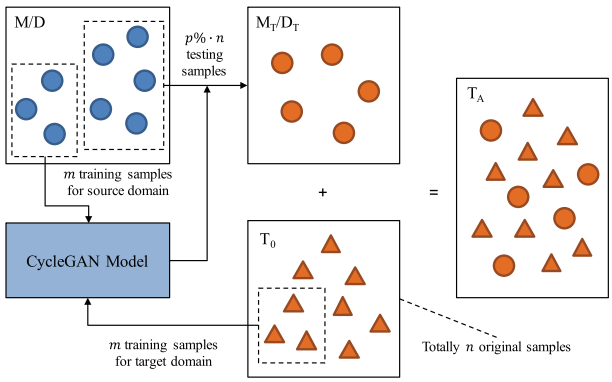


Fig. 4. The proposed adversarial domain adaptation schemes: modality and cross-dataset (M/D) domain adaptation (MDA/DDA). m samples are randomly selected from the source domain and synthesized to the target domain using cycleGAN. Later $p\% \cdot n$ samples are randomly selected to generate the synthesized images and then be combined with n original samples from the target domain. At last, the combination data form the training set.

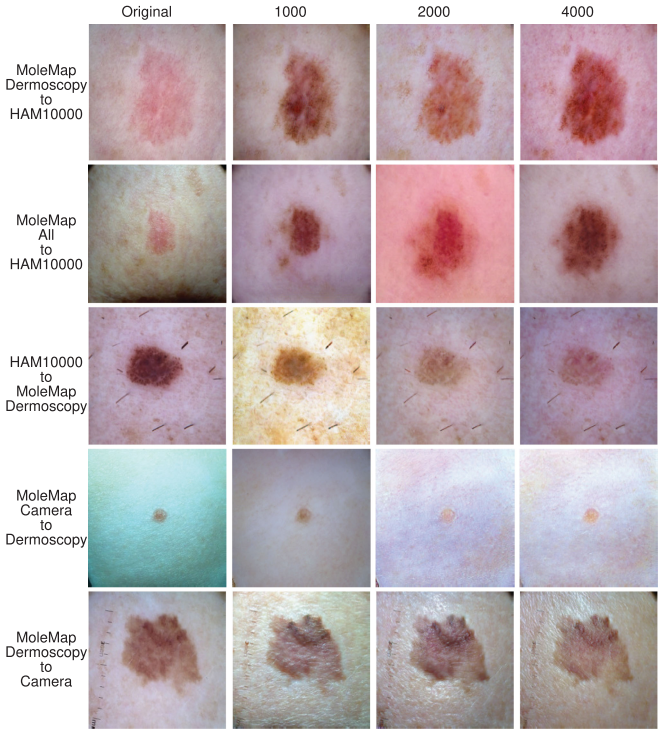


Fig. 5. Illustration of CycleGAN generated images for domain adaptation based on different adaptation model and training epochs.

to both datasets, we propose to use cycleGAN to translate images from MoleMap to HAM (vice versa). The generator G takes a lesion image in MoleMap as input and is trained to produce a new lesion image adapted to the HAM features. The discriminator D is trained to distinguish real HAM images from synthetically generated "fake" images from MoleMap.

We generated synthetic HAM lesion images and introduced variability by adjusting label distributions (benign and cancer ratio). Sample images are shown in Fig. 5. Although higher contrast and light consistency along the lesion boundary is observed in the HAM dataset, the synthesized MoleMap-to-HAM images appear reasonably realistic. Given a sufficiently large training set like MoleMap that covers a large variety of modalities, this early evidence shows that using GAN based model to translate and generate realistic synthetic images is feasible.

We can interpret the benefit of using cycleGAN process from the view of image augmentation. Traditional image augmentation options include rotation, random cropping and mirroring etc. Our cycleGAN method does not conflict with these options and can be used as an addition. Moreover, we can control and have various input options for GAN based methods such as label perturbation and distribution control (for example in some scenarios we prefer to generate more cancer cases). This attribute is useful in medical applications because the medical datasets normally have long-tailed distribution, as the number of different disease types is unbalanced during collection in the real world. Additionally, cycleGAN only requires two unpaired image domains to train the translation model while most other are image-to-image based translation techniques. In dermatology semantics, image-to-image based model need two images to represent healthy and unhealthy status of one unique lesion, which is more difficult to prepare.

2) Modality Domain Adaptation: CycleGAN method not only can be used for domain adaptation between various datasets, but also helps to learn inner domain modality adaptation. As described in Section III, MoleMap dataset consists of two major modalities, dermoscopy images and clinical images. Sample images are shown in the first row of Fig. 1. The camera-to-dermoscopy generator is able to remove the reflection and illumination from the camera taken lesion images, as shown in Fig. 5. According to the empirical results in Section V, this method may be extended to other inner modality adaptation tasks such as magnified vs. unmagnified and polarized vs. unpolarized images.

D. Learning With Synthetic Sets

We trained the skin cancer classification model with original images from the target dataset and transformed synthetic images from the source dataset. We randomly selected $p\%$ of n images from $\mathcal{X}_S$ as auxiliary data. Note that equal proportions were set for each class to maintain the same disease distribution for a fair performance on the testing set. The selected images were passed to the pre-trained source-to-target cycleGAN model for creating the target domain augmentation $\mathcal{X}_T^*$ with similar data distribution and attribute. The combined dataset $\mathcal{X}_T + \mathcal{X}_T^*$ was used to train the model. The ratio of the synthetic images can be controlled under cycleGAN framework. There are two major tasks in evaluating the effectiveness of the domain adapted images. The first is multi-class skin cancer classification on both MoleMap and HAM datasets. The second task is aggregating multiple disease scores and turning them into binary classification such as melanoma-vs-benign and cancer-vs-non cancer.

V. Experiments and Results

In this section, we introduce the system implementation, evaluation metrics and the performance of the proposed methods

TABLE IV
RESULTS ON HAM (H) AS TARGET DATASET WITH PARAMETER BASED TRANSFER LEARNING

Datasets	AC	Accuracy for each class							Melanoma vs Benign			Cancer vs Non-cancer		
		AKIEC	BCC	BKL	DF	MEL	NV	VASC	SEN	SPC	AUC	SEN	SPC	AUC
M (7 cls)	0.707	0.820	0.603	0.755	0.694	0.310	0.718	0.000	0.310	0.980	0.910	0.533	0.902	0.843
H	0.731	0.224	0.519	0.383	0.000	0.228	0.924	0.786	0.228	0.977	0.824	0.343	0.938	0.820
M $\rightarrow$ **H**	0.805	0.296	0.747	0.581	0.735	0.378	0.941	0.929	0.378	0.964	0.887	0.513	0.947	0.897
I $\rightarrow$ **H**	0.907	0.786	0.896	0.830	0.824	0.664	0.967	0.929	0.664	0.980	0.962	0.747	0.971	0.965
I $\rightarrow$ **M** $\rightarrow$ **H**	**0.909**	0.735	0.883	0.827	0.676	0.697	0.971	0.929	0.697	0.982	**0.965**	0.764	0.974	**0.969**

TABLE V
MODEL PARAMETERS

Datasets	Lr_0	Df	ND_E
H & **M**	0.0001	0.1	30
M$\rightarrow$**H** & **I**$\rightarrow$**H**	0.001	0.1	30
I$\rightarrow$**M**$\rightarrow$**H**	0.0001	0.2	15

introduced in Section IV. More supporting experiments are provided in the Appendix.

A. System Implementation

The proposed methods were implemented with Pytorch library on an NVIDIA TITAN X GPU. The implementation codes will be available online at https://github.com/heugyy/Domain-adaptation-Melanoma. We used ResNet 152 [22] as the backbone for classification network, which was trained using an ADAM [44]optimizer. As for CycleGAN, we use the same architecture and training procedure as in [25]. The initial learning rate was Lr_0 with a decay factor of Df for every ND_E epochs in a total of 50 epochs. Since two datasets are in different sizes, different parameter settings $PrS = [Lr_0, Df, ND_E]$ were used and details of parameters are given in the following sub-sections. Before training, all images were normalized to the same scale by dividing the standard deviation (Std) after deducting the mean (Mn), i.e. $\mathcal{X} = (\mathcal{X} - Mn)/Std$, and augmented by horizontal flipping and randomly cropping with the size of 224×224. For all testing data, images were resized to 256×256 and center-cropped to 224×224. Then we normalized the data with zero mean and unit variance before fitting the data into the networks. 30% of HAM images in each class were randomly selected for testing while the others were used for training. 15% of Molemap images in each category were randomly partitioned for testing, while the others were used for training. We choose different percentage of testing samples because MoleMap is a much larger dataset than HAM. The overlapping categories as those in HAM were extracted from MoleMap dataset in all two-dataset related experiments. All the testing sets remained unchanged in all experiment settings.

B. Evaluation Metrics

Four evaluation criteria were utilized to measure the performance of both binary and multi-class classifications for parameter based transfer learning and synthesizing domain adaptation. They are overall accuracy (AC), sensitivity (SEN), specificity (SPC), and AUC (Area under ROC curve) score. **Accuracy** is the overall rate of correctly identified samples. It can be used for both binary classification and multi-class classification. All three other metrics were used for binary classification only. In binary classification, we treated melanoma and cancer conditions as positive, and benign and non-cancer as negative. **Sensitivity** measures the portion of correctly identified positive samples among all positive samples. **Specificity** measures the portion of correctly identified negative samples in all negative samples. These three criteria are defined as:

$$AC = \frac{N_{tp} + N_{tn}}{N}, SEN = \frac{N_{tp}}{N_p}, SPC = \frac{N_{tn}}{N_n} \quad (8)$$

where $N_{tp}, N_{tn}, N, N_p, N_n$ denote number of true positive, true negative, total testing samples, positively classified samples, and negatively classified samples, respectively. In addition, SEN and SPC are different under different classification thresholds. Therefore, AUC is commonly used for the overall measurements, where ROC (receiver operating characteristic curve) is a curve plotting SEN vs SPC under different thresholds.

C. Performance of Parameter Based Transfer Learning

Here, we present the experiments on parameter based transfer learning. We used ImageNet (I) [24] or MoleMap (M) [16] as the source domain, and small dataset HAM (H) [20] as the target domain. As mentioned before, ResNet152 [22] was applied as the deep CNN architecture. Table IV presents both binary and multi-classification results for training M and H from scratch (upper part in Table IV), one-step transfer learning from M to H and from I to H (middle part in Table IV), as well as progressive transfer learning from I to M to H (lower part in Table IV). Following Section V Subsection V-A, parameter settings PrS for the parameter based transfer learning experiments are shown in Table V.

Training from scratch vs one-step transfer learning: It is observed that compared with training from random initialization for the **H** model, improvement of testing results on target dataset is achieved by applying parameters of the pre-trained model that are trained on either dataset MoleMap (**M**) or ImageNet (**I**). This proves that by applying pre-trained knowledge from either large irrelevant dataset or task-same dataset with different probability distribution can enhance the discrimination and generalization ability of the target model. In terms of overall accuracy AC of multi-classification, the results of one-step transfer learning models **M**$\rightarrow$**H** and **I**$\rightarrow$**H** are 7% and 18% higher than training HAM from scratch (**H**) respectively. As for binary classification

TABLE VI
VERIFICATION OF PROGRESSIVE TRANSFER LEARNING

Datasets	AC	Accuracy for each class							Melanoma vs Benign			Cancer vs Non-cancer		
		AKIEC	BCC	BKL	DF	MEL	NV	VASC	SEN	SPC	AUC	SEN	SPC	AUC
C	0.731	-	-	-	-	-	-	-	-	-	-	-	-	-
C→H	0.795	0.459	0.753	0.486	0.059	0.369	0.948	0.786	0.369	0.970	0.884	0.511	0.944	0.884
M→C→H 150E	**0.829**	**0.531**	**0.786**	**0.638**	0.618	**0.484**	0.938	**0.857**	0.484	0.966	**0.906**	0.589	0.948	**0.913**
M→C→H 50E	0.790	0.367	0.760	0.547	0.088	0.414	0.928	0.762	0.414	0.950	0.876	0.550	0.927	0.883
C→M→H	0.803	0.449	0.701	0.498	**0.647**	0.378	**0.951**	0.786	0.378	0.971	0.875	0.501	0.950	0.884

tasks such as melanoma vs benign and cancer vs non-cancer, AUC of one-step transfer learning are better than those of the **H** model. This indicates that one-step transfer learning helps to find a better local optimum for the **H** model than random initialization, which brings the specific benefit of correctly categorizing melanoma and cancer.

Progressive transfer learning vs one-step transfer learning: It is observed that the progressive transfer learning **I→M→H** method outperforms all the other methods in all the evaluation criteria. One-step transfer learning **I→H** ranks the second, slightly lower than **I→M→H**. However, the performance of one-step transfer learning **M→H** is not closely comparable with the other approaches, which is about 10% lower than **I→H**. Although ImageNet is a task-different dataset compared with the task-same dataset MoleMap (M), transferring pretrained knowledge from ImageNet shows more improvement than that from MoleMap. This could be due to three reasons. Firstly, for the source models in both one-step transfer learning experiments, the **I** model has a better local optimum to start with for model **H**. ImageNet is a large dataset in terms of both number of images and categories, but the MoleMap dataset is 10 times smaller than ImageNet in terms of training samples. Secondly, the low-level features from pre-trained ImageNet model are more diverse than those of MoleMap, thereby bringing more discrimination ability for training the target domain model. The last reason might be the influence of MoleMap's inherent data noise, such as incorrect labels (not all images are biopsy verified) and the unclean background of the images, which lowers the overall performance.

Direction comparison of progressive transfer learning: Another set of progressive transfer learning is added as a supporting experimental set, as shown in Table VI. This is performed for comparing the effectiveness of progressive transfer learning with one step transfer learning on the same training depth, as well as comparing the difference between transferring directions.

A reverse direction of progressive learning, namely from a large skin dataset to a natural scene dataset, then to the small skin dataset (**M→C→H**) is compared with the forward direction (**C→M→H**). Instead of using ImageNet, we use CIFAR100 as the source domain representing the natural scene, because the number of images from ImageNet may overwhile the model trained from a relatively small dataset such as Molemap or HAM. The training parameters $PrS = [Lr_0, Df, ND_E]$ for all the transfer learning experiments are $[0.0001, 0.2, 15]$. ResNet is trained from scratch on CIFAR100 (**C**) for 150 iteration epochs, where an accuracy of 73.1% is achieved. Then the pre-trained CIFAR100 model is fine-tuned on HAM10000 (**C→H**) as the baseline. Lastly, the two directions of progressive transfer learning are compared against each other in the last three rows from Table VI. Two sets of **M→C** are iterated with different epochs, i.e. 50 and 150, in the backward direction progressive transfer learning **M→C→H**, achieving accuracies of 79.0% and 82.9% respectively. The last row **C→M→H** shows the results of forward direction, with **C** model trained from scratch with 150 iteration epochs.

From the results shown in Table VI, we have three interesting observations. **Firstly**, experiments using CIFAR100 as the source domain is more difficult to train on the MoleMap dataset and achieves lower accuracy compared with using ImageNet model as the source. This may be caused by the intrinsic property of CIFAR100 model whose size and category distribution is much smaller than ImageNet and the deep layer semantic information is not as sufficient to learn as ImageNet for skin disease classification. **Secondly**, the best overall accuracy is achieved when training on the intermediate dataset CIFAR100 with 150 iteration epochs of **M→C→H**. Comparing with 1-step transferring learning (**C→H**), the pretrained model of large natural scene dataset (CIFAR100) works better on small skin disease dataset (HAM) with pretrained knowledge of skin disease (Molemap), which further validates our previous assumption that progressive transfer learning works better than 1-step transfer learning. **Thirdly**, if the iteration epochs of backward direction progressive transfer learning **M→C→H** remain the same as forward direction **C→M→H**, the performance becomes worse. This validates the importance of the order of progressive transfer learning, i.e. transferring knowledge from large natural scene data to large amounts of skin disease data, and then to the target small skin disease dataset.

All in all, from the experimental results, we can see that it achieves better results when iterating larger epochs and exploring deeper semantic knowledge in progressive transfer learning backwards, but deeper means more computational expensive. Comparing with the large computing resources consumed by very deep networks, progressive transfer learning improves the classification performance with the limited one-step transfer learning depth.

D. Qualitative Analysis of Adversarial Domain Adaptation

In order to validate the performance of synthesizing domain adaptation, CycleGAN was run between different dataset

groups, as shown in Fig. 5. The left column of the image indicates the cycleGAN source domain and target domain. The experiments are divided into two groups, dataset-different group and modality-different group, which is in accordance with the settings in Section V-E. We first qualitatively validate the influence of number of training samples on the final output. CycleGAN was trained over various training numbers N, i.e. $N = 1000, 2000, 4000$, for both source and target domains. Another important hyper-parameter for GAN is the number of training iterations (epoch). For 1000, 2000, and 4000 training samples, the corresponding numbers of epochs were 100, 200 and 200, respectively.

For dataset domain adaptation (DDA), both HAM alike images and MoleMap alike images were generated. HAM alike images were generated from two protocols, MoleMap dermoscopic modality only (the 1st row in Fig. 5), and MoleMap all modalities (the 2nd row in Fig. 5). It can be seen that the center lesion of the synthetic HAM images are enhanced by various extents. In the first row, the contours of lesions are roughly the same in regarding to different validation hyper parameters N, while in the second row the contours are enlarged or zoomed as the number of training samples turns bigger. Meanwhile, all images in the second row show a blurred background after translation, this may attributes to the clinical images included in the source domain. Synthesis of MoleMap dermoscopy alike images ((the 3rd row in Fig. 5)) are generated by training with HAM dermoscopic images as source domain. The synthetic images are prone to be weakened in terms of lesion boundaries and textures, which show strong similarity to the MoleMap style dermoscopic images.

For modality domain adaptation (MDA), the two modalities of the MoleMap dataset were transformed alternatively. 4th row in Fig. 5 presents results of normal camera images to dermoscopic images, it can be observed that the colors of the fake dermoscopic images are tuned similar to the target domain, and the textures are blurred in terms of background and contours. As for the dermoscopic modality to normal camera translation, synthetic images contain richer details and higher contrast in textures, the boundaries become sharper and illumination reflections are generated around the contours, compared with the source dermoscopic images.

Furthermore, Fig. 6 illustrates CycleGAN results in terms of less dependant on source and target domain. The images in the same row shows rarely changed basic structures, which is constrained by the cycle-consistent loss. However, due to the less dependency between the source and the target domain, the appearance of the region of interest is completely different and may cause label noise during training, as shown in Fig. 6. Blue rectangles indicate the potential Region of Interest (RoI) of original images while green one indicate the RoI of the synthesized images. In the synthesized images, either the those regions of interest are shrunk in size (the first row in Fig. 6) or the non-interested regions are highlighted (the second row in Fig. 6). Therefore, during dataset domain adaptation, the synthesized images are selected randomly from those images with little bioinformatics noise, such as eyes, lips and nose etc.

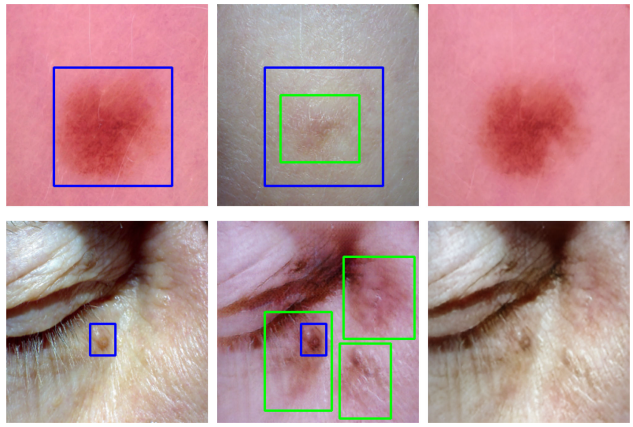


Fig. 6. This figure shows the cycleGAN synthesizing images from less dependent target domain. The columns from left to right show original images, synthesized images and images that revert back to the source domain, respectively. The first row shows the HAM dermoscopy images while the second row shows the Molemap camera images.

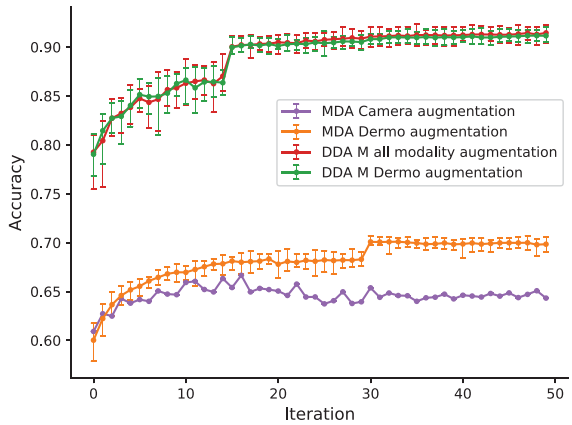


Fig. 7. Overall accuracy of multi-classification over epochs of modality domain adaptation (MDA) and dataset domain adaptation (DDA). We found that MDA dermoscopic images are more helpful than clinical camera images for data augmentation (orange vs purple) on MoleMap dataset. Same trend can also be observed for DDA based experiments on HAM dataset.

E. Quantitative Analysis of Dataset and Modality Adversarial Domain Adaptation

We conduct multi-class skin classification experiments on the MoleMap and HAM datasets with synthesizing augmentation to evaluate the performance of adversarial domain adaptation. As introduced in Section III, both the categories and modality constitution between HAM and MoleMap are different. In order to align the categories of two datasets, the training and testing samples in MoleMap were selected from the same categories as HAM. We first perform a set of comprehensive experiments to validate how augmentation ratio and number of training samples will affect the multi-class skin disease classification, as shown in Fig. 7. More specifically, we train a set of cycleGAN models with training sample number $N = 1000, 2000, 4000$ in both domains and use those generated images to augment the target domain training set by the percentage of 10%, 20%, 50%, 100%.

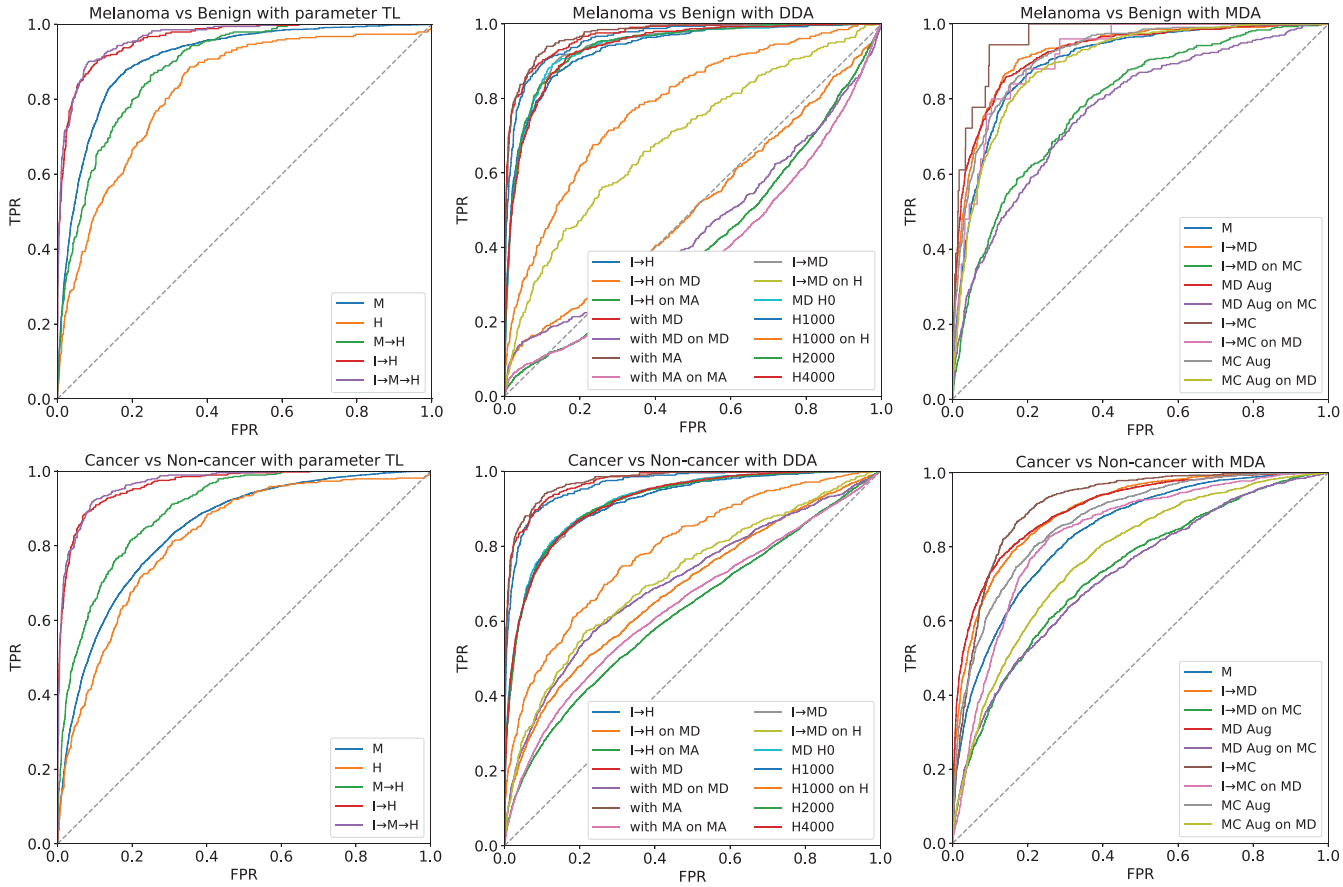


Fig. 8. This figure introduces ROC curves of melanoma vs benign and cancer vs. non-cancer on parameter based transfer learning (TL), data domain adaptation (DDA) and modality domain adaptation (MDA) models.

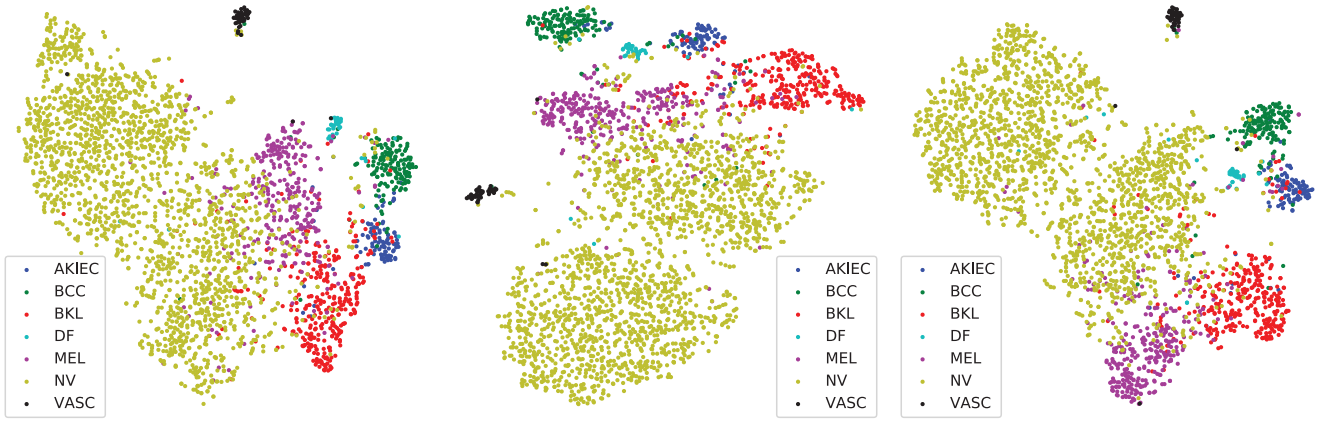
TABLE VII
DATASET DOMAIN ADAPTATION (DDA) USING HAM DATASET AS TARGET DOMAIN AND MOLEMAP AS SOURCE DOMAIN

Datasets	AC	Accuracy for each class							Melanoma vs Benign			Cancer vs Non-cancer		
		AKIEC	BCC	BKL	DF	MEL	NV	VASC	SEN	SPC	AUC	SEN	SPC	AUC
I→H	0.907	0.786	0.896	0.830	0.824	0.664	0.967	0.929	0.664	0.980	0.962	0.747	0.971	0.965
← on MD	0.395	0.201	0.514	0.447	0.424	0.093	0.927	0.000	0.093	0.977	0.509	0.424	0.855	0.684
← on MA	0.310	0.117	0.410	0.352	0.199	0.038	0.920	0.000	0.038	0.976	0.411	0.329	0.854	0.623
I→H+MD*	0.920	0.786	0.883	**0.848**	**0.824**	0.751	0.970	**0.976**	0.751	0.982	0.969	0.803	0.975	0.972
← on MD	0.442	0.243	0.587	0.532	0.505	0.115	0.872	0.000	0.115	0.966	0.454	0.495	0.819	0.703
I→H+MA*	**0.923**	**0.857**	**0.903**	0.833	0.706	**0.778**	**0.971**	0.929	0.778	0.981	**0.973**	0.823	0.974	**0.975**
← on MA	0.341	0.146	0.511	0.349	0.149	0.069	0.914	0.000	0.069	0.963	0.388	0.424	0.802	0.640

These settings lead to 12 sets of results for each run and in total there are 4 runs, where each run with different augmentation rate and cycleGAN model is represented as a curve of a different color in Fig. 7. Here, 50 training epochs are iterated as default for every combination set. The overall accuracy of multi-class classification are used for evaluation. On each curve, the dots show the average of the overall accuracy while the error bar indicates the minimum and maximum overall accuracy of 12 sets over 50 iteration epochs.

In the following two sub-sections, we explain the experiments of dataset domain adaptation (DDA) and modality domain adaptation (MDA) in details, where the best performance of the 12 sets in each run, namely the maximum error bar in Fig. 7, represents for the according domain adaptation result. All relevant AUC-ROC experimental results are shown in Fig. 8.

Dataset Domain Adaptation (DDA): Task on *HAM Augmentation* follows the same settings as described in Section V-E, and results are shown in Table VII. The target domain is the HAM dataset which are dermatoscopic images, while the source domains are MoleMap containing both modality images (MA) and the MoleMap dermoscopy (MD) respectively. We used parameters $PrS = [0.0001, 0.2, 15]$ to train both MD and MA augmented HAM dataset. Both cycleGAN transformed augmentations outperform the ImageNet fine-tuning results **I→H**



(a) t-SNE of **I→H** model (b) t-SNE of **I→H+MD*** model (c) t-SNE of **I→H+MA*** model

Fig. 9. This figure shows the embedding distribution comparison of dataset domain adaptation (DDA) results.

(the first row in Table VII). The augmented MA training samples (**I→H+MA***) consist of half dermoscopy and half clinical images from MoleMap, and it achieves the best performance in terms of AC and AUC compared to **I→H** and dermoscopic only augmented model **I→H+MD***.[4] Meanwhile, in order to validate the generalization capability of the proposed domain adaptation scheme, we conduct extra experiments on the trained model back to its source domain dataset. (← **on MD**) in the Table VII indicates that we test the trained target model back to its MoleMap dermoscopic domain. It can be observed that all adversarial augmented model **I→H+MA*** and **I→H+MD*** show an increase, when compared to the performance of **I→H** model. This experiment proves that the dataset domain adaptation with adversarial learning improves both the model's classification performance and generalization capability between different datasets.

In order to verify and visualize the proposed model in the statistical manner, t-Distributed Stochastic Neighbor Embedding (t-SNE) is employed for the three main models of HAM augmented dataset domain adaptation, as shown in Fig. 9. These three models are finetuning ImageNet pretrained model on HAM, **I→H** (Fig. 9(a)), finetuning ImageNet pretrained model on **MD*** augmented HAM, **I→H+MD*** (Fig. 9(b)) and finetuning ImageNet pretrained model on **MA*** augmented HAM, **I→H+MA*** (Fig. 9(c)) respectively. From the figure, it is shown that the points around the boundaries are more crowded and blurry in Fig. 9(a) compared with the other two sub-figures, especially in the adjacent area between MEL, BKL and DF class. This shows dataset domain adaptation (DDA) improves the discrimination capacity of the classification model, and the improvement is significant for some diseases, such as melanoma and DF.

[4]The best performance **I→H+MD*** as source augmentation is 10% augmentation with cycleGAN 2000 training model. For **MA** as source, the best performance relies on 10% augmentation with cycleGAN 1000 training results, which are listed in the first row of the middle part and bottom part in Table VII.

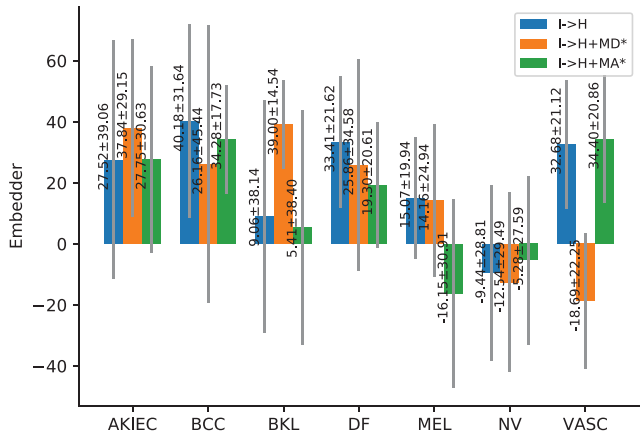


Fig. 10. This figure shows the mean and standard deviation of embeders of different models by disease classes.

TABLE VIII
INTRA-CLASS DISTANCE EVALUATION

Model	**I→H**	**I→H+MD***	**I→H+MA***
Intra-class embeder mean	52.30	52.73	53.20

To further quantitatively analyze the t-SNE in Fig. 9, inter-class and intra-class variation is calculated and shown in Fig. 10 and Table VIII, respectively. In Fig. 10, each color group represents the embedding mean and standard deviation of the model, where each rectangle bar is the mean and the line bar is the deviation. The smaller deviation means that inter-class data distances are less variant, or denser. As shown in Fig. 10, **I→H+MD*** model has the least variance in the classes AKIEC and BKL among all three models, while **I→H+MA*** model has the least variances in the classes AKIEC, BCC, DF, NV and VASC. As for Table VIII, the 2-dimension t-SNE mean center for each class is calculated firstly before calculating the distance mean of every two-center pair. The mean evaluates the intra-class variance, where the higher value means that data of different

TABLE IX
USING A DIFFERENT NETWORK TO TESTIFY OUR ASSUMPTION. EXPERIMENTS OF TABLE VII WITH DENSENET

Datasets	AC	Accuracy for each class							Melanoma vs Benign			Cancer vs Non-cancer		
		AKIEC	BCC	BKL	DF	MEL	NV	VASC	SEN	SPC	AUC	SEN	SPC	AUC
I→H	0.906	0.745	0.864	0.833	**0.824**	0.703	0.964	**0.905**	0.703	0.978	0.967	0.766	0.971	0.969
← on MD	0.418	0.244	0.546	0.481	0.435	0.102	0.865	0.000	0.102	0.978	0.582	0.450	0.840	0.696
← on MA	0.344	0.141	0.439	0.448	0.221	0.084	0.871	0.000	0.084	0.954	0.511	0.379	0.837	0.654
I→H+MD*	0.910	0.745	0.870	0.833	0.765	**0.730**	**0.967**	0.881	0.730	0.983	0.971	0.784	0.973	0.972
← on MD	0.433	0.235	0.581	0.524	0.457	0.110	0.865	0.000	0.110	0.981	0.604	0.472	0.816	0.697
I→H+MA*	**0.914**	**0.786**	**0.877**	**0.878**	0.735	0.715	0.966	**0.905**	0.715	0.983	**0.974**	0.774	0.976	**0.974**
← on MA	0.344	0.119	0.505	0.408	0.204	0.079	0.908	0.000	0.079	0.966	0.459	0.418	0.813	0.645

TABLE X
DATASET DOMAIN ADAPTATION (DDA) USING MOLEMAP AS TARGET DOMAIN AND HAM AS SOURCE DOMAIN

Datasets	AC	Accuracy for each class							Melanoma vs Benign			Cancer vs Non-cancer		
		AKIEC	BCC	BKL	DF	MEL	NV	VASC	SEN	SPC	AUC	SEN	SPC	AUC
I→MD	0.806	0.853	0.809	**0.854**	0.821	0.614	0.687	0.000	0.614	0.967	**0.941**	0.762	0.902	0.919
← on H	0.521	0.439	0.734	0.708	0.677	0.604	0.473	0.000	0.604	0.675	0.692	0.717	0.586	0.718
I→MD+H	**0.814**	0.857	**0.816**	0.847	0.810	0.571	**0.787**	0.000	0.571	0.973	0.940	0.755	0.913	**0.922**
I→H+MD* (H1000)	0.810	0.859	0.803	0.846	**0.837**	**0.645**	0.712	0.000	0.645	0.966	0.934	0.768	0.904	0.915
I→H+MD* (H2000)	0.810	**0.876**	0.781	0.843	**0.837**	0.601	0.732	0.000	0.601	0.969	0.940	0.739	0.915	0.918
I→H+MD* (H4000)	0.807	0.859	0.806	0.849	0.832	0.535	0.754	0.000	0.535	0.976	0.938	0.738	0.915	0.919
← on H	0.555	0.561	0.818	0.693	0.735	0.766	0.485	0.000	0.766	0.660	0.782	0.823	0.571	0.791

TABLE XI
PARAMETERS FOR DATASET DOMAIN ADAPTATION METHOD

Datasets	Lr_0	Df	ND_E
M	0.001	0.1	30
I→MD & **I→MC**	0.0001	0.2	15
I→H+MD* & **I→H+MC***	0.0001	0.1	30

TABLE XII
MODALITY DOMAIN ADAPTATION (MDA) USING MOLEMAP AS TARGET DOMAIN AND SOURCE DOMAIN

Dataset	AC	Melanoma vs Benign			Cancer vs Non-cancer		
		SEN	SPC	AUC	SEN	SPC	AUC
M (25 cls)	0.606	0.320	0.975	0.900	0.506	0.906	0.839
I→MD	0.689	0.402	0.981	**0.925**	0.657	0.916	0.899
← on MC	0.457	0.187	0.972	0.790	0.336	0.910	0.732
I→H+MD*	**0.707**	0.546	0.972	0.924	0.725	0.899	**0.904**
← on MC	0.472	0.224	0.968	0.771	0.443	0.857	0.725
I→MC	0.666	0.111	0.999	**0.962**	0.591	0.932	**0.914**
← on MD	0.510	0.000	0.999	0.915	0.458	0.900	0.835
I→H+MC*	**0.667**	0.516	0.960	0.918	0.615	0.899	0.873
← on MD	0.512	0.646	0.903	0.895	0.560	0.817	0.776

classes are more variant and separate, which in turn brings better classification results. The **I→H+MA*** model has the highest mean value, which demonstrates its best classification performance among these three models.

Extra experiments with DenseNet: To validate the generalization ability of the proposed Dataset Domain Adaptation (DDA) method that is effective for various network architectures. Extra results in Table IX show the accordant performance compared with the ResNet model showing in Table VII. The DenseNet results in Table IX show the similar trend as ReseNet in Table VII. Training on both sources of dataset domain adaptation augmented HAM data (**I→H+MA*** and **I→H+MD***) outperforms training on the original HAM dataset (**I→H**), and **I→H+MA*** performs the best. Synthesized image as source achieves the best performance in terms of overall accuracy and AUC and most of single class accuracy, although there is a slight drop in terms of DF accuracy.

Experiments with *MoleMap Augmentation* is shown in Table X. The experiments are based on the parameters $PrS = [0.0001, 0.1, 30]$. Same as the last *HAM Augmentation* experiments, we present the results of ImageNet transferred model **I→MD** and adversarial augmented model **I→H+MD***. Because MoleMap is a much larger dataset than HAM in terms of number of images per category, it is difficult to maintain the same augmentation ratio as 10%, 20%, 50%, 100%. Therefore, we decide to transform all HAM data to MD for augmentation and report various cycleGAN models' $N = 1000, 2000, 4000$ results. It is surprising to find that models with various N don't differ that much compared to the parameter transferred model. We have the hypothesis that the number of augmented images from HAM are not enough to change the model training that much, and HAM dermoscopic images are relatively unified compared with MoleMap dataset where images are acquired from a much diverse sources and cohorts. However, we are still able to figure out that when the trained model are tested on the source dataset, the adversarial learning augmented models enjoy more generalization capabilities. Furthermore, we also conduct an experiment to see whether original HAM combined with original MD training samples lead to better performance. From **MD + H**, we observe that it achieves the best multi-class accuracy across all models. This may seem contradictory to our hypothesis that adversarial augmented samples should help train a better performing model than un-transformed images. The

TABLE XIII
COMPARISON RESULTS OF EACH DISEASE ON MODALITY DOMAIN ADAPTATION (MDA)

Datasets	AC	Accuracy for each class												
		0	1	2	3	4	5	6	7	8	9	10	11	12
M (25 cls)	0.606	0.035	0.617	0.518	0.000	0.000	0.643	0.727	0.221	0.061	0.033	0.067	0.000	0.387
I→MD	0.689	0.188	**0.692**	0.609	0.071	0.000	**0.429**	**0.833**	0.387	0.364	0.207	**0.529**	0.353	0.353
← on MC	0.457	0.040	0.320	0.186	0.000	0.000	0.286	0.629	0.015	0.074	0.058	0.000	0.000	0.000
I→H+MD*	**0.707**	**0.250**	0.639	**0.629**	0.071	0.000	0.357	0.762	**0.467**	**0.416**	**0.262**	0.471	**0.412**	**0.588**
← on MC	0.472	0.040	0.292	0.237	0.000	0.000	0.214	0.600	0.323	0.194	0.101	0.077	0.077	0.143
I→MC	0.666	**0.080**	0.154	0.343	0.091	0.000	0.429	0.829	0.246	**0.056**	**0.130**	**0.308**	**0.357**	**0.690**
← on MD	0.510	0.094	0.177	0.113	0.000	0.000	0.143	0.619	0.013	0.000	0.067	0.118	0.059	0.362
I→H+MC*	**0.667**	0.000	**0.532**	**0.524**	0.091	0.000	0.429	**0.857**	**0.369**	0.046	0.007	0.231	0.077	0.143
← on MD	0.512	0.000	0.393	0.056	0.000	0.000	0.286	0.524	0.080	0.000	0.000	0.118	0.000	0.177

13	14	15	16	17	18	19	20	21	22	23	24
0.598	0.760	0.158	0.378	0.320	0.023	0.580	0.139	0.129	0.859	0.355	0.704
0.696	0.739	0.469	0.592	0.402	0.160	0.738	0.410	0.316	**0.832**	**0.604**	**0.818**
0.552	0.111	0.160	0.171	0.187	0.000	0.336	0.126	0.008	0.827	0.309	0.511
0.768	**0.821**	**0.500**	**0.653**	**0.546**	**0.360**	**0.783**	**0.472**	**0.368**	0.826	0.590	0.795
0.638	0.272	0.120	0.220	0.224	0.111	0.446	0.231	0.084	0.771	0.282	0.497
0.759	0.400	0.488	0.424	0.111	**0.729**	0.642	**0.328**	**0.839**	0.484	**0.776**	0.483
0.288	0.500	0.061	0.542	0.000	0.603	0.536	0.151	0.682	0.407	0.609	0.085
0.741	**0.741**	**0.520**	**0.439**	**0.516**	0.056	**0.648**	0.294	0.513	**0.893**	0.527	**0.767**
0.464	0.348	0.406	0.082	0.646	0.040	0.458	0.272	0.211	0.799	0.500	0.563

reason might be that unlike HAM images are pathologically verified, tele-dermatology labelled MoleMap dataset contain a certain ratio of noise. Many lesions and atypical nevus are overly conservative labelled as cancer to avoid false negative decisions. Therefore, un-transformed images may serve a strong regularization for the MoleMap target trainined model and lead to better performance.

Modality Domain Adaptation (MDA): To verify the performance of various models on modality domain adaptation, we conduct experiments using MoleMap dermoscopy (MD) and MoleMap camera images (MC) for full 25 skin disease classification. Following Section V Subsection V-A, parameter settings PrS for the parameter based transfer learning experiments are shown in Table XI. Here we produce two separate sets of experiments on MD and MC, each setting contains one-step parameter based transfer model (**I** $\rightarrow$ **MD** and **I** $\rightarrow$ **MC**) as baseline and our proposed adversarial domain adopted technique. Same as dataset domain adaptation (DDA) experiments, trained models are cross-tested on the source domain to verify the generalization capability.[5] Table XII shows the overall results and the corresponding accuracy for each disease class is shown in Table XIII.

The results show that MD augmentation (**I→H+MD***) achieves the best overall accuracy and cancer vs non-cancer AUC, while fine-tuning ImageNet without augmentation slightly outperforms it by 0.01% for melanoma vs benign classification. This is due to the reason that MD images provide rich and noise-free textual details for classification. When MC images are adversarial transformed into MD, the cycleGAN model can simply zoom in and remove artifacts from the background to generate good quality auxiliary training samples for MD. What is more important to be aware of is that in Table XIII, the most lethal cancer class of melanoma has received a large margin improved detection rate when trained with the model **I→H+MD***. It further demonstrates the effectiveness of our proposed method on melanoma detection.

[5]The rows **I→H+MD*** and **I→H+MC*** have the best result based on 10% augmentation with 4000 training samples in cycleGAN. For all other augmentation combination of the target dataset, results are illustrated as the orange line in Fig. 7.

VI. DISCUSSION AND CONCLUSION

Recent years have witnessed the blossom of deep convolutional neural networks in medical image analysis which requires numerous data for training to get a robust model. These normally come from multiple datasets of different sources for the same task. However, different datasets cannot be unified directly as a technique of training data augmentation, since the bias and variance of different datasets often lead to drop of performance. As far as we know, there are not any previous works that quantify the combination variance of different inner sub-category datasets in skin disease classification. In this paper, we investigate the difference of classification performance by transferring and adapting knowledge from different skin cancer datasets. We also study whether generative domain adaptation is capable of being an effective dataset augmentation technique for datasets bias. Our experimental results lead to the following observations:

1) Initializing deep networks with parameters of the pre-trained model trained on task-same but constitution-different larger datasets improves the model discrimination ability. Therefore, we can see when applying progressive transfer learning outperforms one-step transfer learning.
2) The experiments show that cycleGAN domain adaptation helps to alleviate the domain shift between two

TABLE XIV
NOTATION SUMMARY OF THE TABLES IN SECTION V

Notation	Meaning
MD*/MA*/MC*	Synthesized images with MD/MA/MC as source domain
M (7 cls)	Training MoleMap(of 7 classes) from scratch and test on MoleMap(7 classes)
H	Training HAM from scratch and test on HAM
M → H	Using pretrained MoleMap training from scratch model to fine tune on HAM and test on HAM
I → H	Using pretrained ImageNet model to fine tune on HAM and test on HAM
I → M→ H	Using pretrained ImageNet model to fine tune on MoleMap and then fine tune on HAM and test on HAM
← on MD	Using the model on the above row to test on MD
← on MA	Using the model on the above row to test on MA
I → H+MD*	Using pretrained ImageNet model to fine tune on MD* augmented HAM and test on HAM
I → H+MA*	Using pretrained ImageNet model to fine tune on MA* augmented HAM and test on HAM
I → MD	Using pretrained ImageNet model to fine tune on MD and test on MD
← on H	Using the model on the above row to test on H
I → MD+H	Using pretrained ImageNet model to fine tune on the combination of original MD and H and test on MD
I → H+MD*(H#)	Using pretrained ImageNet model to fine tune on HAM augmented MD(cycleGAN is trained with # images) and test on MD
M (25 cls)	Training MoleMap(of 25 classes) from scratch and test on MoleMap of 25 classes
I → MD	Using pretrained ImageNet model to fine tune on MD of 25 classes and test on MD of 25 classes
← on MC	Using the model on the above row to test on MC of 25 classes
I → H+MD*	Using pretrained ImageNet model to fine tune on MD* augmented MC of 25 classes and test on MC of 25 classes
I → MC	Using pretrained ImageNet model to fine tune on MC of 25 classes and test on MC of 25 classes
← on MD	Using the model on the above row to test on MD of 25 classes
I → H+MC*	Using pretrained ImageNet model to fine tune on MC* augmented MD of 25 classes and test on MD of 25 classes

different datasets and improve the performance for the models with its augmented samples. However, we need to aware that careful selection of augmentation rate is needed. Although cycleGAN is performed to minimize such distribution shift, there is still variance caused by other factors such as labelling noise can not be recovered.

3) The overall accuracy improvement of dataset domain adaptation (DDA) augmentation is better than modality domain adaptation (MDA) augmentation. This may rely on the intrinsic property and limitation of cycleGAN domain adaptation. This limitation becomes dramatic if the source domain differs largely from the target domain. The visual differences between different modalities (MA vs MD) are larger than dermascopic images among two datasets.
4) For categories of importance, especially melanoma, the correctly classified rate has been greatly improved when applying adversarial domain adaptation. This may be caused by melanoma in MoleMap has less labelling noise (clinicians tend to be more careful about this label).
5) The adversarial domain adaptation methods improve the model generalization. The performance drop among two different sources as shown in Table I, VII, X, XIII, XII, has been alleviated through dataset domain adaptation (DDA). For modality domain adaptation (MDA), similar trends have been observed. This might be caused by the pre-trained knowledge that is inherited from the source domain. Although the source domain images are generated into the target domain alike images, they still keep original knowledge to some extends, such as the textural and pathological information. This will in turn improves the model's discrimination and generalization ability testing on the source domain.

In this work, we quantitatively validate the model generalization for different datasets from two perspectives. One is applying parameter-based progressive transfer learning to share transferable knowledge from task-different source domain and task-same but dataset-different intermediate domain with target domain. In the second method, we generalize the model for different datasets by integrating images from other datasets after translating with cycle consistent generative networks (CycleGAN). In this way, the model can be generalized for dataset-different domain as well as modality-different domain. Our experiments show the improvements for both overall multi-class classification accuracy and binary classification accuracy on both source domain datasets and target domain datasets. The improvement in binary classification is especially outstanding, which, in real cases, is more expected as the missing rate of melanoma shall be lowered. In the future, to further improve the classification performance, an algorithm can be developed that may contain discriminant features from both the training sets. The fusion could be developed by constructing a hybrid training parameter set from the two training set parameters which were extracted on individual data sets. Although this work applies domain adaptation to skin disease imaging dataset augmentation, we believe this scheme may inspire more studies on applications that lack training data, especially in general medical imaging applications.

APPENDIX

In the Appendix, the notations in the experiment tables in Section V are summarized, as shown in Table XIV.

REFERENCES

[1] V. Gray-Schopfer, C. Wellbrock, and R. Marais, "Melanoma biology and new targeted therapy," *Nature*, vol. 445, no. 7130, pp. 851–857, 2007.

[2] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: Cancer J. Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[3] T. Mendonca, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira, "Ph2 - A dermoscopic image database for research and benchmarking," in *Proc. Int. Eng. Med. Biol.*, 2013, pp. 5437–5440.

[4] Q. Abbas, M. E. Celebi, and I. F. Garcia, "Hair removal methods: A comparative study for dermoscopy images," *Biomed. Signal Process. Control*, vol. 6, no. 4, pp. 395–404, 2010.

[5] P. Schmid, "Segmentation of digitized dermatoscopic images by two-dimensional color clustering," *IEEE Trans. Med. Imag.*, vol. 18, no. 2, pp. 164–171, Feb. 1999.

[6] J. Yang, X. Sun, L. Jie, and R. Paul, "Clinical skin lesion diagnosis using representations inspired by dermatologist criteria," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, vol. 11, pp. 1258–1266.

[7] N. Codella, J. Cai, M. Abedini, R. Garnavi, A. Halpern, and J. R. Smith, "Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, Oct. 2015, pp. 118–126.

[8] Y. Gu, J. Zhou, and B. Qian, "Melanoma detection based on Mahalanobis distance learning and constrained graph regularized nonnegative matrix factorization," in *Proc. Winter Conf. Appl. Comput. Vision*, Mar. 2017, pp. 797–805.

[9] Y. Rivenson, Z. Gorocs, H. Gunaydin, Y. Zhang, H. Wang, and A. Ozcan, "Deep learning microscopy," *Optica*, vol. 4, no. 11, pp. 437–1443, 2017.

[10] P. Rajpurkar *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning," 2017. [Online]. Available: https://arxiv.org/abs/1711.05225v3

[11] M. E. Celebi, Q. Wen, H. Iyatomi, K. Shimizu, H. Zhou, and G. Schaefer, "A state-of-the-art survey on lesion border detection in dermoscopy images," *Dermoscopy Image Anal.*, 2015, pp. 97–129.

[12] N. C. Codella *et al.*, "Deep learning ensembles for melanoma recognition in dermoscopy images," *IBM J. Res. Develop.*, vol. 61, no. 4, pp. 5.1–5.15, 2017.

[13] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[14] L. Yu, H. Chen, Q. Dou, J. Qin, and P.-A. Heng, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Trans. Med. Imag.*, vol. 36, no. 4, pp. 994–1004, Apr. 2017.

[15] F. Xie, H. Fan, Y. Li, Z. Jiang, R. Meng, and A. Bovik, "Melanoma classification on dermoscopy images using a neural network ensemble model," *IEEE Trans. Med. Imag.*, vol. 36, no. 3, pp. 849–858, Mar. 2017.

[16] Z. Ge, S. Demyanov, R. Chakravorty, A. Bowling, and R. Garnavi, "Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images," in *Proc. Conf. Med. Image Comput. Comput. Assisted Intervention*, 2017, pp. 250–258.

[17] A. C. F. Barata, E. M. Celebi, and J. Marques, "A survey of feature extraction in dermoscopy image analysis of skin cancer," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 3, pp. 1096–1109, May 2019.

[18] S. Pathan, K. G. Prabhu, and P. Siddalingaswamy, "Techniques and algorithms for computer aided diagnosis of pigmented skin lesions–A review," *Biomed. Signal Process. Control*, vol. 39, pp. 237–262, 2018.

[19] D. Gutman *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC)," in *Proc. Int. Symp. Biomedical Imag.*, Apr. 2018, pp. 168–172.

[20] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, 2018, Art. no. 180161. doi: 10.1038/sdata.2018.161.

[21] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Conf. Comput. Vision Pattern Recognit.*, Jun. 2016, pp. 770–778.

[23] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. Conf. Comput. Vision Pattern Recognit.*, Miami, FL, USA, 2009, pp. 248–255.

[25] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. Int. Conf. Comput. Vision*, 2017, pp. 2223–2232.

[26] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, Oct. 2018.

[27] G. Csurka, "Domain adaptation for visual applications: A comprehensive survey," *arxiv.org/abs/1702.05374*, 2017.

[28] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[29] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Proc. Conf. Neural Inf. Process. Syst.*, 2016, pp. 343–351.

[30] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. Conf. Comput. Vision Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 5967–5976.

[31] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh, "Recycle-GAN: Unsupervised video retargeting," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 119–135.

[32] A. Rahman Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *Proc. Neural Inf. Process. Syst. Workshop Deep Learn. Speech Recognit. Related Appl.*, 2009, vol. 1, no. 9, p. 39.

[33] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proc. Int. Conf. Mach. Learn. Workshop Unsupervised Transfer Learn.*, 2012, pp. 17–36.

[34] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks," in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.

[35] G. Wang *et al.*, "Interactive medical image segmentation using deep learning with image-specific fine tuning," *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1562–1573, Jul. 2018.

[36] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2018, vol. 1, pp. 328–339.

[37] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. Comput. Vision Pattern Recognit.*, 2014, pp. 1717–1724.

[38] B. Chu, V. Madhavan, O. Beijbom, J. Hoffman, and T. Darrell, "Best practices for fine-tuning visual classifiers to new domains," in *Proc. Eur. Conf. Comput. Vision (TASK-CV)*, 2016, pp. 435–442.

[39] M. Everingham, L. V. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst.*, 2012, vol. 25, pp. 1097–1105.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[42] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. CVPR*, 2017, vol. 1, no. 2, pp. 4700–4708.

[43] T. Zhou, P. Krahenbuhl, M. Aubry, Q. Huang, and A. A. Efros., "Learning dense correspondence via 3d-guided cycle consistency," in *Proc. Comput. Vision Pattern Recognit.*, 2016, pp. 117–126.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: https://arxiv.org/abs/1412.6980