scientific reports

Check for updates

OPEN

Investigating the key principles in two-step heterogeneous transfer learning for early laryngeal cancer identification

Xinyi Fang[1,2], Chak Fong Chong[1,2], Kei Long Wong[1,3], Marco Simões[2] & Benjamin K. Ng[1]✉

Data scarcity in medical images makes transfer learning a common approach in computer-aided diagnosis. Some disease classification tasks can rely on large homogeneous public datasets to train the transferred model, while others cannot, i.e., endoscopic laryngeal cancer image identification. Distinguished from most current works, this work pioneers exploring a two-step heterogeneous transfer learning (THTL) framework for laryngeal cancer identification and summarizing the fundamental principles for the intermediate domain selection. For heterogeneity and clear vascular representation, diabetic retinopathy images were chosen as THTL's intermediate domain. The experiment results reveal two vital principles in intermediate domain selection for future studies: 1) the size of the intermediate domain is not a sufficient condition to improve the transfer learning performance; 2) even distinct vascular features in the intermediate domain do not guarantee improved performance in the target domain. We observe that radial vascular patterns benefit benign classification, whereas twisted and tangled patterns align more with malignant classification. Additionally, to compensate for the absence of twisted patterns in the intermediate domains, we propose the Step-Wise Fine-Tuning (SWFT) technique, guided by the Layer Class Activate Map (LayerCAM) visualization result, getting 20.4% accuracy increases compared to accuracy from THTL's, even higher than fine-tune all layers.

Laryngeal cancer is one of the most frequently reported cancers in the head and neck. According to Deng et al.[1], by 2017, instances of laryngeal cancer have increased by 58.6% in 27 years worldwide. At the same time, as the mortality rate increases, the average age of the disease slips. Although the final diagnosis of laryngeal lesions relies on clinical biopsy results, such as histopathological images of laryngeal squamous cell carcinoma, studies[2–4] have focused on improving the classification performance and interpretability of these images. However, different types and stages of lesions can also be reflected in the vascular structure changes of the larynx[5]. In otolaryngology endoscopy, the emerging narrow-band image (NBI) technique is favored over traditional white light imaging in the diagnosis and early detection of lesions[6] since it keeps the blue and the green light with specific wavelengths, which can better show the morphology of the mucosal epithelium and the epithelial vascular. With the development of deep learning, combining deep learning techniques and NBI for computer-aided diagnosis, it is possible to differentiate different kinds of lesions and obtain results that are extremely close to pathological diagnosis[7].

The scale of medical image datasets is typically small, making it impractical to train deep neural models since those models are data-eager. One possible solution to overcome this limitation is to utilize transfer learning. Transfer learning refers to inheriting the knowledge acquired from the source domain to assist the tasks in the target domain, such as the classification task. Applying transfer learning in image analysis is ubiquitous by combining the Artificial Neural Network (ANN) and reusing partial model parameters pre-trained on the source domain, commonly a larger dataset scale, such as ImageNet[8].

Most transfer learning applied to medical image diagnosis is one-step transfer learning, meaning it directly uses the pre-trained model from the source domain and then fine-tunes it in the target domain. However, Tan et al.[9] pointed out that the performance of one-step transfer learning would be lower than expected if the features between the source domain and target domain were highly distinct. Matsoukas et al.[10] proved that the

[1]Faculty of Applied Sciences, Macao Polytechnic University, Macao 999078, China. [2]Department of Informatics Engineering, Centre for Informatics and Systems of the University of Coimbra, University of Coimbra, Coimbra 3000, Portugal. [3]Department of Computer Science and Engineering, University of Bologna, Bologna 40100, Italy. ✉email: bng@mpu.edu.mo

features from ImageNet have little reuse on the chest x-ray and lymph node-stained sections. Raghu et al.[11] also concluded that ImageNet as the source domain does not significantly improve the performance of diabetic retina diagnosis and chest x-ray detection.

One possible solution might be using an intermediate domain that works as a bridge between the source and target domains in a two-step transfer learning process, providing more target-like features for the model to learn. However, studies about two-step transfer learning remain rare, especially for two-step heterogeneous transfer learning and selecting an appropriate intermediate domain.

Existing studies show that successful two-step transfer learning could rely on selecting a homogeneous intermediate domain with semantically identical images and tasks to the target domain. However, the lack of homogeneous and large-scale public datasets to serve as intermediate domains for laryngeal blood vessels makes it hard for models to extract features that support the classification tasks in the target domain. Thus, it is crucial to explore the semantically nonidentical intermediate domain, such as using another organ's images or other type of blood vessel, for the laryngeal blood vessel classification task.

Therefore, we make two hypotheses to explore the feasibility of using the proposed two-step heterogeneous transfer learning in laryngeal cancer identification and summarising fundamental principles for intermediate domain selection. ***Hypothesis (1)***: ImageNet is an appropriate source domain for laryngeal blood vessel classification. ***Hypothesis (2)***: performing two-step heterogeneous transfer learning in laryngeal blood vessel classification by using other organ vessels as the intermediate domain is feasible.

The main contributions of this article are: This is the first work to investigate the effectiveness of two-step heterogeneous transfer learning on laryngeal blood vessel classification task using the color fundus photographies of Diabetic Retinopathy[12] as the intermediate domain. The diabetic retinopathy imaging domain is chosen for its larger size and clear depiction of vascular structures, aligning with the assumptions of the proposed THTL framework. Combining Layer Class Activate Map (LayerCAM)[13] and ResNet50, we visualize the model's decision-making process and conclude a key principle for selecting the intermediate domain: radial blood vessels favor benign classifications. In contrast, twisted and tangled vessels link to malignancy. We propose an advanced fine-tuning strategy called Step-Wise Fine-Tuning (SWFT) to improve the classification performance in THTL with the aid of LayerCAM. 11 deep learning models are applied in this work, and we have summarised the best-performing model for four classification scenarios.

Related work

Transfer learning

A formal definition of transfer learning is given by Pan and Yang[14], considering a source domain $\mathscr{D}_S = \{\mathscr{X}_S, P(X_S)\}$ and a target domain $\mathscr{D}_T = \{\mathscr{X}_T, P(X_T)\}$, where the $\mathscr{X}$ and $P(X)$ are the feature space and marginal data distribution of the corresponding domain. Given a source domain learning task $\mathscr{T}_S$ and a target domain learning task $\mathscr{T}_T$, where the task corresponding to each domain contains that domain's label space $\mathscr{Y}$ and predictive function $f(\cdot)$. Using knowledge learned from the $\mathscr{D}_S$ and $\mathscr{T}_S$, transfer learning aims to improve the performance of the target predictive function $f_T(\cdot)$ in the $\mathscr{D}_T$. Notice that neither domains nor tasks are the same.

Transfer learning can be further categorized into homogeneous and heterogeneous based on the difference between the source and target domain. Homogeneous transfer learning strictly restricts the feature space, and the label space of the source and target domain must be the same. Failing to meet any of these conditions is a heterogeneous transfer learning[15]. Moreover, since our method utilizes distinct domains and addresses different tasks, the proposed THTL can be categorized as cross-domain transfer learning[16,17].

In this work, we propose to investigate the effectiveness of two-step heterogeneous transfer learning for laryngeal blood vessel classification by given the definition of the intermediate domain $\mathscr{D}_I = \{\mathscr{X}_I, P(X_I)\}$, the learning task of the intermediate domain $\mathscr{T}_I = \{\mathscr{Y}_I, f_I(\cdot)\}$, and a strict condition, which is $\mathscr{X}_S \neq \mathscr{X}_I \neq \mathscr{X}_T$, $\mathscr{Y}_S \neq \mathscr{Y}_I \neq \mathscr{Y}_T$, and $\mathscr{T}_S \neq \mathscr{T}_I \neq \mathscr{T}_T$.

One-step transfer learning

In medical image classification tasks, directly transferring the models' parameters from ImageNet to the target domain is a conventional approach and usually obtains a well-performed result, such as some works of laryngeal lesions classification[18,19]. However, they lack of verifying the effectiveness of using ImageNet as the source domain for transfer learning.

Using ImageNet as the source domain might not guarantee good results. Heker and Greenspan[20] pointed out that, in some circumstances, using ImageNet as the source domain might not bring an expected performance under medical image classification tasks. Still, there is a chasm between the medical and natural images. Alzubaidi et al.[21] proved that the breast cancer classification task will perform better if the source domain transferred is close to the target domain, not ImageNet. Xie and Richmond[22] verified that, for the chest x-ray classification task, models pre-trained on gray-scale ImageNet exceed those pre-trained on original ImageNet.

To the best of the authors' knowledge, it is an open question whether using ImageNet as a source domain for classifying laryngeal blood vessels is effective.

Two-step heterogeneous transfer learning

We utilize the concept of sequential transfer learning[23] in natural language processing to medical image classification, as manifested in the knowledge that can be transferred from the source domain to the intermediate domain first, then transferring the knowledge learned from the intermediate domain to the target domain.

De Matos et al.[24] obtained the improved results in breast cancer histopathologic image classification. They used ImageNet as the source domain to get the pre-trained models, then trained a model in an intermediate domain with histopathologic images to acquire the knowledge to exclude the blank space of a breast tumor

histopathologic image in the target domain. Alkhaleefah et al.[25] compared the classification performance of one-step and two-step heterogeneous transfer learning in mammogram images and concluded that two-step heterogeneous transfer learning outperforms one-step's. Notice that the source domain was ImageNet, and the datasets in the intermediate and target domains were mammogram images. The paper of Alzubaidi et al.[26] demonstrated that the classification performance of diabetic foot ulcers was enhanced by transferring the knowledge learned from the skin cancer to the feet' skin image first, then transferring the knowledge to the diabetic foot ulcers. Meng et al.[27] applied two-step transfer learning to detect COVID-19. They used ImageNet as the source domain, which was heterogeneous to the tuberculosis (TB) CT image in the intermediate domain, making the model learn the features close to the target domain. The target domain was COVID-19 CT images. This two-step transfer learning improved the overall accuracy by 1.36%.

Despite the entities above being successful, there are limitations to two-step heterogeneous transfer learning. Whether in terms of imaging modality or semantics, the datasets in the intermediate domains for those studies are very close to those in the target domains, even homogeneous. This is feasible for medical image types that have received much attention. However, it is challenging to find a publicly available large-scale dataset as the intermediate domain for two-step heterogeneous transfer learning for relatively less discussed medical image types, such as laryngeal blood vessels. Therefore, it is critical to seek an alternative intermediate domain that might be formed by different imaging modalities and semantically nonidentical to the laryngeal blood vessel.

Deep-learning models applied in medical image classification

In this work, different types of models will be employed to verify the feasibility of two hypotheses we proposed. Morid et al.[28] analysed the frequently used deep learning models in medical image analysis from 2012 to 2020. In particular, Inception[29], ResNet[30], and Visual Geometry Group (VGG)[31] were commonly used in endoscopic images, therefore, we select InceptionV3[32], ResNet18, ResNet50, and VGG19 in our experiments. We also include DenseNet[33], which was commonly practiced in lung studies, into consideration. The selection for DenseNet is DenseNet121 and DenseNet169. Additionally, the lightweight model MobileNet V2[34] is efficient and performed excellently in soft tissue classification[35]. EfficientNetB0 made its debut in laryngeal disease classification[36], with the highest accuracy and relatively low memory consumption, proving the value of EfficientNet in the classification of laryngeal diseases. Additionally, the Vision Transformer (ViT)[37] has shown great potential in laryngeal lesion classification, such as in probe-based confocal laser endomicroscopy (pCLE)[38], as well as in its variant, ViT-based model for laryngeal histopathology images[2,4]. Thus, we also include MobileNet V2, and four state-of-art models (the EfficientNetV2-S from EfficientNetV2[39], ViT-B/16 from Vision Transformer, SwinV2-T from Swin Transformer V2[40], and MaxVit-T from Multi-Axis Vision Transformer[41]) in our experiments.

Methods

Verification of Hypothesis (1)

Expt.#1 and Expt.#2 aim to investigate the effectiveness of using ImageNet as the source domain in larynx cancer classification by comparing the performance difference between models with and without pre-training on ImageNet. The designs are illustrated in Figure 1. Models in Expt.#1 are pre-trained on ImageNet and all their parameters are updated based on the target domain. These models are marked with **. On the contrary, models in Expt.#2 start with the random initial weights and only use the CE-NBI dataset for training, validating, and testing. They are marked with ^.

Verification of Hypothesis (2)

Expt.#3 and Expt.#4 aim to investigate the feasibility of THTL in larynx cancer classification tasks by comparing the performance difference between one-step and two-step heterogeneous transfer learning. The illustration is shown in Figure 2.

Expt.#3 investigates the performance of one-step heterogeneous transfer learning. ImageNet is used as the source domain. CE-NBI dataset is used as the target domain. ImageNet pre-trained models, ResNet18, ResNet50, MobileNet V2, Inception_V3, Densenet121, Densenet169, VGG19, ViT-B/16, EfficientNetV2-S, SwinV2-T, and MaxVit-T are used. The pre-trained weights are kept as the initial weights of models. To assess the impact of

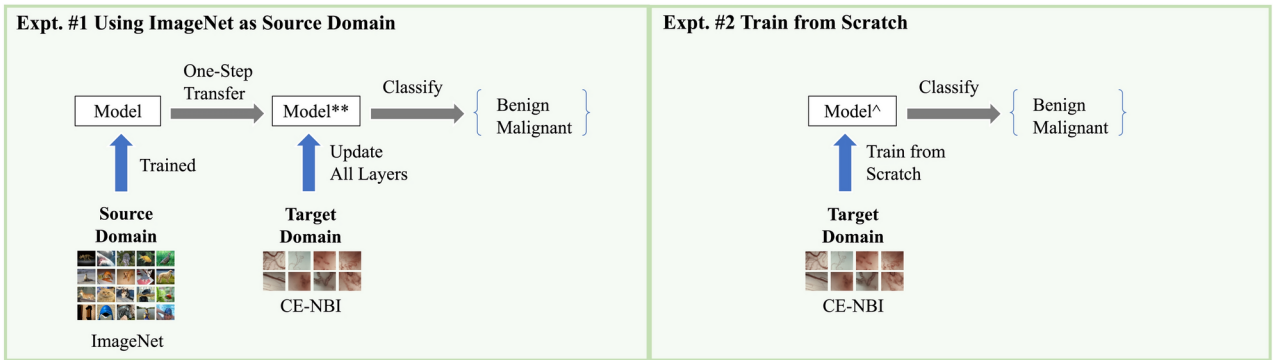


Fig. 1. Experiment design for hypothesis (1).

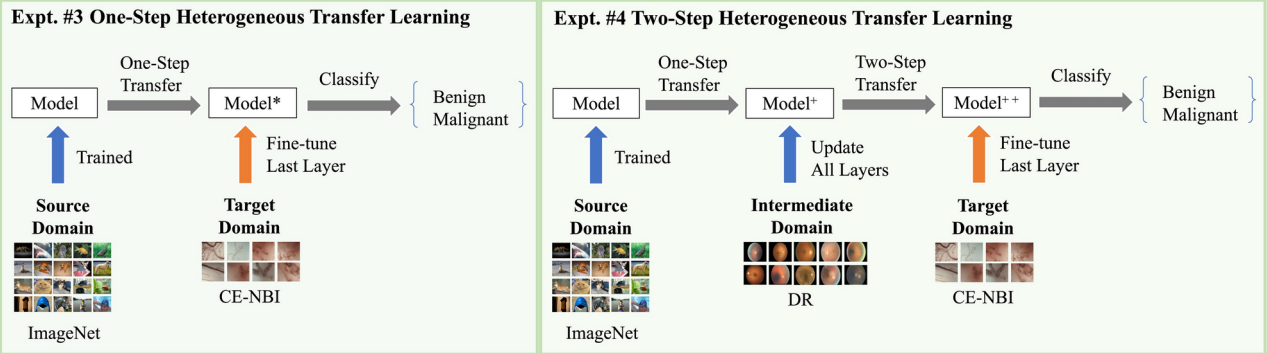


Fig. 2. Experiment design for hypothesis (2).

domain transfer on the target domain, we only fine-tune the fully connected (fc) layer of the model based on the target domain and modify the number of output classes to two (benign and malignant). The fine-tuned models are marked with an asterisk *.

Expt.#4 investigates the performance of two-step heterogeneous transfer learning. The source domain is ImageNet. Since we aim to capture the features of the intermediate domain to provide to the target domain, these models are fully re-trained in the intermediate domain using diabetic retinopathy images. Particularly, our objective extends beyond solely classifying the diabetic retina. We aim to transfer the knowledge of distinguishing blood vessels to the target domain task using the models trained in the intermediate domain. Consequently, we choose not to adjust the hyperparameters specifically for the models. To remain consistent, we utilize the hyperparameters listed in Details of Experiment Configurations for all experiments. These fully-trained models are marked with +. After that, we transfer model+ and only fine-tune the model's last layer on the target domain to assess the impact of intermediate domain transfer on the target domain. In this phase, the fine-tuned models are marked with ++, and then they are used to do the laryngeal blood vessel classification task.

Datasets and data pre-processing methods

In this section, three public datasets, the data preparation and pre-processing work, as well as the evaluation metrics for classification performance will be introduced.

ImageNet (source domain)

In this paper, ImageNet is used as the source domain task for experiments. ImageNet is the most frequently used, well-known dataset for pre-trained models in the image field, the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012) of ImageNet[42]. This dataset contains 1,000 categories of natural images over one million. The weights of pre-trained models based on ImageNet are used as the initial weights for later training.

CE-NBI dataset (target domain)

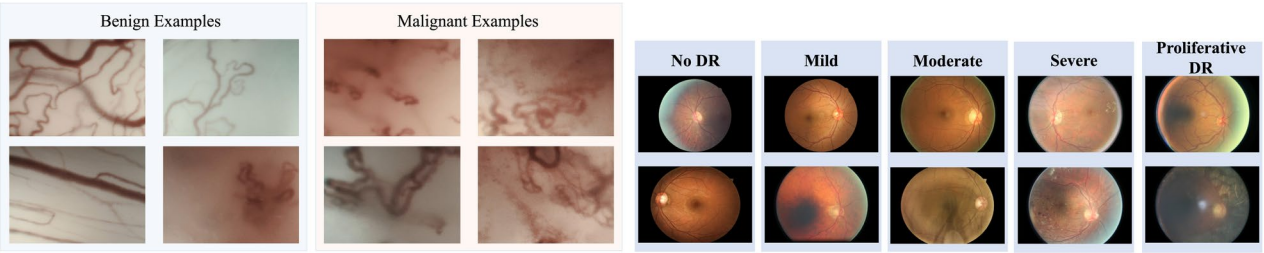
The Contact Endoscopy combined with Narrow Band Imaging (CE-NBI) was released by Esmaeili et al.[43] in 2019. This dataset is used as the target domain task for laryngeal blood vessel classification. In this dataset, the subepithelial blood vessels of the vocal folds are enhanced and magnified. The dataset contains 11,144 images of vocal fold subepithelial blood vessels, categorized into benign and malignant according to the type of lesions. Depending on laryngeal histopathology diagnostics, both benign and malignant are further subdivided into different lesions. The benign category includes Amyloidosis, Cyst, Granuloma, and other eight kinds of diseases; the malignant are subdivided into squamous cell carcinoma (SCC), high grade dysplasia, and Carcinoma in situ. As the target domain task of our research, to better avoid the class imbalance issue, we use the lesion type label to classify the CE-NBI images into benign and malignant classes. The samples of this dataset are shown below in Figure 3. The first two rows are the samples of the benign class, and the last two are the malignant class samples.

In this dataset. the images are randomly divided as the ratio of 7 : 2 : 1 into training, validation, and testing set, using random seed 123. Thus, there are 7,803 images for training, 2,228 for validation, and 1,113 for testing. Detailed data volume statistics can be found in Table 1.

Also, we apply data augmentation techniques during the training phase by resizing the images to meet the model's requirement (e.g., $384 * 384$ for EfficientNetV2-S, $299 * 299$ for InceptionV3, $256 * 256$ for SwinV2-T, and $224 * 224$ for the rest models) and randomly horizontally flipping the images. For the validation and testing phases, we only resize the images. Moreover, image normalization is applied in all three phases, using $mean = [0.485, 0.456, 0.406]$ and standard deviation equals [0.229, 0.224, 0.225] .

DR dataset (intermediate domain)

We aim to utilize a modality and semantically nonidentical dataset to the target domain dataset, but under the premise that they are both images of blood vessels. A dataset of Diabetic Retinopathy (DR) Detection published on Kaggle[12] is used as the intermediate dataset. The modality of this dataset is digital color fundus photography, which is different from the CE-NBI endoscopy image. This dataset serves to identify the severity of DR by color fundus photographies into five classes: no sign of Diabetic Retinopathy, a mild symptom of DR, a moderate



(a) Samples of CE-NBI dataset. **(b)** Samples of DR dataset.

Fig. 3. Samples for the target and intermediate domains.

Class	Training Set	Validation Set	Testing Set	Total
Benign	5,361	1,531	765	7,657
Malignant	2,442	697	348	3,487

Table 1. Data volume statistics of CE-NBI dataset.

	No DR	Mild	Moderate	Severe	Proliferative DR
Training Set	18,067	1,710	3,704	611	495
Validation Set	7,743	733	1,588	262	213
Total	25,810	2,443	5,292	873	708

Table 2. Data volume statistics of diabetic retinopathy detection dataset.

symptom of DR, a severe symptom of DR, and Proliferative DR. The task is also different to the CE-NBI dataset. Therefore, the DR dataset satisfies the restrictions of the intermediate domain for two-step heterogeneous transfer learning for blood vessel classification.

We only use the training set of this public dataset since the competition is closed and released test dataset are not labelled. Therefore, we treat the training set as the whole dataset (DR dataset), and further split it into training and validation set for our experiment. We use the training set to train the models and select the model with the highest performance on the validation set. Thus, the testing set is optional. The DR dataset contains 35,126 images, randomly divided into a training and a validation set by the ratio of $7:3$, with the random seed 123. Thus, the training set contains 24,587 images, and the validation set contains 10,539 images. Samples of the DR dataset can be found in Figure 3b. Two samples of the current class are in each column. Table 2 shows each class's detailed data volume statistics.

The data augmentation configuration follows the same operations as the CE-NBI dataset.

Details of experiment configurations

The experiments are conducted using Python programming language, with all deep learning models developed using the PyTorch framework[44]. The NVIDIA RTX A6000 and NVIDIA GeForce RTX 2080 Ti GPUs are utilized for the training, validation, and testing phases. We apply the following configurations to every experiment: including experiments #1, #2, #3, and #4. The batch size is set to 32, with a maximum training epoch of 70 and implementation of an early stopping technique. The parameters for the early stopping technique are set to a patience of 10 and delta of 0. The Adam optimizer[45] is utilized, with learning rate decay implemented through a learning rate step scheduler with $step\_size = 5$ and $gamma = 0.5$. The initial learning rate for EfficientNetV2-S, VGG19, SwinV2-T, and MaxVit-T is set to $1e-4$, while for the other models it is set to $1e-3$. The loss function employed is Cross Entropy. The input size of the image varies depending on the model. For EfficientNetV2-S, the input size is adjusted to $384 * 384$. For InceptionV3, the input size is set to $299 * 299$. For SwinV2-T, the input size is set to $256 * 256$. For other models, the input size is standardized at $224 * 224$. To avoid the contingency of the experiment, each model is run 3 times and then the average of the results is recorded.

Evaluation metrics

This article uses accuracy, precision, recall, F1-Score, and AUC as the evaluation metrics with the help of confusion matrix. Additionally, we use the macro average (MA) to balance the effects of the classes due to the imbalance in the target domain classes. The range of all matrices is from zero to the positive one, and the closer to 1 means the better the result.

LayerCAM visualization

Class Activation Maps (CAM) can generate images with highlighted regions for convolutional neural networks (CNNs), enabling one to observe the decision-making logic of CNN classification in greater detail. Based on this characteristic, CAM is widely used to enhance the interpretability of medical image classification. For instance, Grad-CAM combined with deep learning models can highlight the critical regions that a model focuses on when classifying histopathological images of oral squamous cell carcinoma[46]. Additionally, volumes of interest (VOIs) identified by Grad-CAM can be extracted for further classification without requiring labeled lymph node regions[47].

Compared to Grad-CAM, LayerCAM has the distinct advantage of generating reliable class activation maps by identifying the most relevant pixels for target objects[48] across multiple layers of a CNN, rather than relying solely on the final convolutional layer. LayerCAM employs gradient-based techniques to emphasize the significance of different locations within the feature map for a specific class and provide more detailed heatmap[13]. By adding these class-specific activation maps to the original input image, LayerCAM produces a visual representation that effectively emphasizes the regions of the image that significantly contribute to the network's decision-making process.

In this work, we resort to the commonly used deep learning model in the image classification field, the ResNet50, to perform the task. ResNet50 contains four extensive basic modules called Layer 1 to Layer 4, in Pytorch. By combining the LayerCAM with the ResNet50, the attention area of the model in each Layer can be represented in the image.

Given a predicted image, the CAM calculated by the LayerCAM can be described using the following equations[13]:

$$M^c = ReLU(\sum_k (w_{ij}^{kc} \cdot A_{ij}^k)) \tag{1}$$

where,

$$w_{ij}^{kc} = ReLU(g_{ij}^{kc}) \tag{2}$$

Specifically, w_{ij}^{kc} represents the weight of the spatial coordinate (i, j) within the k-th feature map, and ^{c} is the class. In our case, ^{c} is the class benign or malignant. By applying the rectified linear unit function (ReLU), the gradients of this coordinate g_{ij}^{kc} remain if it is greater than zero; otherwise, it turns to zero. Then, LayerCAM multiplies the activation value A_{ij}^k of that coordinate with the weight w_{ij}^{kc} to get the class activation map for that certain layer, then linearly combines all channel dimensions to calculate the final class activation map M^c.

Therefore, in our case, when combined with ResNet50, each predicted and labeled image will generate four class activation maps, one for each of the four basic modules (Layer 1 to 4). We choose ResNet50 **, which performs best among four experimental scenarios, as the'standard answer', and compare the decision-making process of ResNet50^{++} in Expt.#4 with it to summarize the principle of intermediate domain selection.

Step-wise fine-tuning method

Since transfer learning is usually combined with deep-learning models or CNNs, several fine-tuning strategies have been adopted for medical image classification. As summarized in Kandel and Castelli[49], there were three common fine-tuning strategies for transfer learning: 1) only fine-tune the fully connected layer and keep entire network weights frozen; 2) fine-tune the entire network weights; 3) fine-tune the layers close to the output while keeping the layers close to input frozen.

Despite those common fine-tuning techniques, some flexible fine-tuning techniques can also bring effectiveness improvements, such as differential evolution based fine-tuning[50] and block-wise fine-tuning[51], which they fine-tuned the last two residual blocks of ResNet18 while keeping other weights frozen. In addition, this work is inspired by layer-wise fine-tuning in Tajbakhsh et al.[52] and Sharma and Mehra[53]. They fine-tuned the smallest unit in the CNN sequentially, such as the convolutional layer and the fully connected layer.

The examination and analysis of convolutional networks have demonstrated that lower layers, positioned closer to the input, primarily specialize in extracting texture-related features. Conversely, higher layers, situated closer to the output, are more class-specific and capture more semantic information, i.e., labels of the images[54,55]. Consequently, implementing our proposed SWFT facilitates the retention of intermediate domain knowledge within the lower-layer parameters of the model while simultaneously enabling the model's adaptation to the target domain in the top-layer parameters.

In this work, we introduce a different fine-tuning strategy called the Step-Wise Fine-Tuning strategy (SWFT) to gradually increase the number of layers fine-tuned from back to front for the model in the second step of THTL. Our proposed methodology is not only distinguished from Tajbakhsh et al.[52] and Sharma and Mehra[53] in terms of the specific model and dataset employed, but we also incorporate the visualization of the LayerCAM as a reference for the number of steps in SWFT.

The design of SWFT is shown in Figure 4. With the ResNet model (ResNet has four major modules, called Layer One to Four), the SWFT can be split into six steps: Step 1 fine-tunes the fully connected layer of the model; Step 2 fine-tunes Layer 4 and the fully connected layer; Step 3 fine-tunes Layer 3, Layer 4, and the fully connected layer. The layers involved in fine-tuning are added one by one up to Step 5. The last step, Step 6, is to update all parameters of the model.

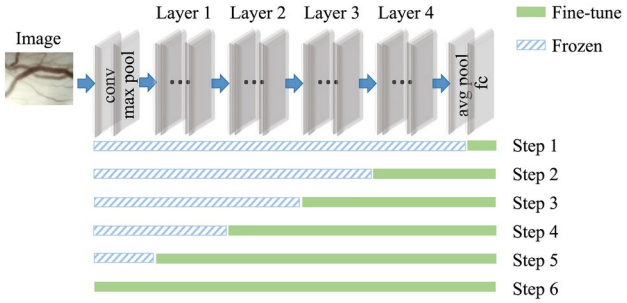


Fig. 4. Demonstration of Step-Wise Fine-Tuning for ResNet.

Expt.	Model	Epoch	Acc.[1]	Precision			Recall			F1-Score		AUC
				B[2]	M[3]	MA[4]	B	M	MA[5]	B	M	
#1	DenseNet121**	32	0.976	0.983	0.961	0.972	0.982	0.964	0.973	0.983	0.962	0.998
	DenseNet169**	30	0.969	0.976	0.954	0.965	0.979	0.946	0.963	0.977	0.950	0.996
	InceptionV3**	27	0.968	0.974	0.954	0.964	0.979	0.943	0.961	0.977	0.948	0.997
	MobileNet V2**	32	0.969	0.979	0.947	0.963	0.976	0.954	0.965	0.977	0.950	0.997
	ResNet18**	28	0.968	0.975	0.951	0.963	0.978	0.945	0.962	0.977	0.948	0.995
	ResNet50**	35	0.958	0.969	0.935	0.952	0.970	0.931	0.951	0.970	0.933	0.993
	VGG19**	25	0.975	0.981	0.963	0.972	0.983	0.959	0.971	0.982	0.961	0.997
	ViT-B/16**	26	0.951	0.964	0.924	0.944	0.966	0.920	0.943	0.965	0.922	0.983
	EfficientNetV2-S**	21	0.980	**0.988**	0.962	0.975	0.983	0.973	0.978	0.985	0.968	**0.999**
	SwinV2-T**	20	0.980	**0.988**	0.962	0.975	0.983	0.973	0.978	0.985	0.968	**0.999**
	MaxVit-T**	18	**0.982**	**0.988**	**0.969**	**0.978**	**0.986**	**0.974**	**0.980**	**0.987**	**0.971**	**0.999**
#2	DenseNet121	68	0.945	0.952	0.928	0.940	0.968	0.894	0.931	0.960	0.910	0.987
	DenseNet169	58	0.948	0.962	0.917	0.939	0.962	0.917	0.939	0.962	0.917	0.987
	InceptionV3	48	**0.964**	**0.976**	**0.938**	**0.957**	**0.972**	**0.946**	**0.959**	**0.974**	**0.942**	**0.993**
	MobileNet V2	42	0.950	0.963	0.920	0.942	0.964	0.920	0.942	0.964	0.920	0.987
	ResNet18	51	0.950	0.959	0.929	0.944	0.968	0.910	0.939	0.964	0.919	0.988
	ResNet50	56	0.947	0.956	0.926	0.941	0.967	0.902	0.935	0.961	0.914	0.984
	VGG19	26	0.936	0.948	0.910	0.929	0.960	0.884	0.922	0.954	0.897	0.974
	ViT-B/16	37	0.925	0.934	0.902	0.918	0.958	0.851	0.904	0.946	0.875	0.972
	EfficientNetV2-S	53	0.944	0.949	0.933	0.941	0.971	0.884	0.928	0.959	0.908	0.984
	SwinV2-T	43	0.860	0.874	0.823	0.849	0.931	0.703	0.817	0.901	0.757	0.925
	MaxVit-T	54	0.940	0.951	0.920	0.935	0.963	0.890	0.926	0.957	0.903	0.984

Table 3. Results of Expt.#1. and Expt.#2. [1] Test Accuracy. [2] Benign class. [3] Malignant class. [4] Macro Avg. for Precision. [5] Macro Avg. for Recall

We have compared the performance of our proposed method with the conventional fine-tuning strategy, which involves fine-tuning either the last fully connected layer or all layers of the model, and an adaptive fine-tuning method called Low-Rank Adaptation (LoRA). The experimental results show that our proposed SWFT achieves comparable accuracy to the best performance in Expt.#1 and surpasses LoRA. Notably, the overall accuracy and malignant recall in Step 3 exceeds those achieved by fine-tuning all layers in the second step in THTL.

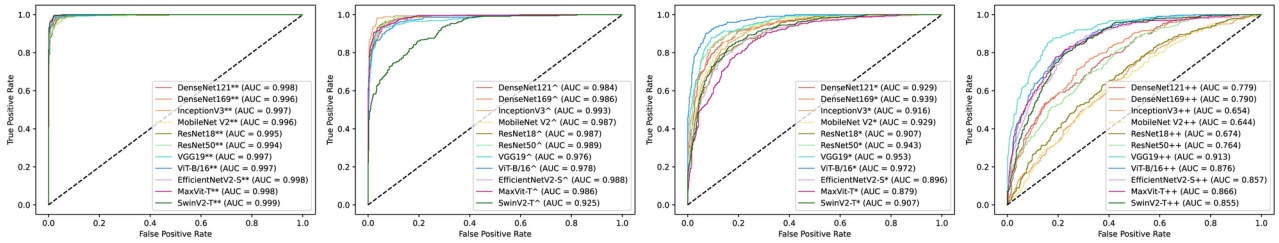
Results and discussions

Results to verify Hypothesis (1)

The results of Expt.#1 and Expt.#2 are concluded in Table 3. And the ROC curves for these two experiments are presented in Figure 5a and Figure 5b. We select the run that is closest to the mean test accuracy presented in Table 3 to generate the ROC curve.

In Expt.#1, ImageNet is used as source domain. All models perform well, among them, the MaxVit-T achieves the best performance (98.2% accuracy) with the least epochs (18). Meanwhile, the ViT-B/16 gets the most negligible result, with an accuracy of 95.1% and AUC of 98.3%. The EfficientNetV2-S and SwinV2-T also perform well in terms of precision for the benign class (98.8%) and AUC (99.9%).

In Expt.#2, ImageNet is not used as source domain, therefore all the models are directly trained in the target domain. The most significant difference compared with the Expt.#1 is the number of epochs. Expt.#2 takes a



(a) ROC curve for Expt.#1 **(b)** ROC curve for Expt.#2 **(c)** ROC curve for Expt.#3 **(d)** ROC curve for Expt.#4

Fig. 5. ROC curves for Experiment One to Four.

Expt.	Model	Epoch	Acc.[1]	Precision			Recall			F1-Score		AUC
				B[2]	M[3]	MA[4]	B	M	MA[5]	B	M	
#3	DenseNet121*	42	0.866	0.883	0.823	0.853	0.929	0.730	0.829	0.905	0.773	0.929
	DenseNet169*	35	0.880	0.907	0.818	0.862	0.920	0.792	0.856	0.913	0.805	0.939
	InceptionV3*	32	0.854	0.873	0.805	0.839	0.922	0.706	0.814	0.897	0.752	0.915
	MobileNet V2*	44	0.856	0.878	0.799	0.839	0.918	0.719	0.818	0.897	0.757	0.928
	ResNet18*	42	0.836	0.858	0.774	8.816	0.911	0.670	0.790	0.884	0.718	0.908
	ResNet50*	55	0.886	0.906	0.837	0.871	0.930	0.788	0.859	0.918	0.812	0.943
	VGG19*	43	0.878	0.907	0.811	0.859	0.915	0.794	0.855	0.911	0.802	0.940
	ViT-B/16*	50	**0.915**	**0.930**	**0.879**	**0.905**	**0.947**	**0.844**	**0.896**	**0.939**	**0.861**	**0.972**
	EfficientNetV2-S*	37	0.820	0.865	0.719	0.792	0.875	0.701	0.788	0.870	0.710	0.879
	SwinV2-T*	70	0.841	0.880	0.753	0.816	0.890	0.734	0.812	0.885	0.743	0.905
	MaxVit-T*	46	0.809	0.852	0.706	0.779	0.874	0.668	0.771	0.863	0.686	0.875
#4	DenseNet121^{++}	26	0.750	0.761	0.691	0.726	0.927	0.360	0.643	0.836	0.473	0.778
	DenseNet169^{++}	31	0.733	0.734	0.724	0.729	0.959	0.236	0.597	0.832	0.355	0.789
	InceptionV3^{++}	25	0.689	0.694	0.544	0.619	**0.982**	0.046	0.514	0.813	0.085	0.657
	MobileNet V2^{++}	23	0.678	0.695	0.420	0.557	0.949	0.083	0.516	0.802	0.138	0.648
	ResNet18^{++}	52	0.689	0.702	0.516	0.609	0.953	0.110	0.532	0.808	0.182	0.674
	ResNet50^{++}	44	0.738	0.743	0.702	0.723	0.946	0.282	0.614	0.832	0.402	0.766
	VGG19^{++}	70	**0.839**	**0.880**	0.746	**0.813**	0.886	**0.736**	**0.811**	**0.883**	**0.741**	**0.913**
	ViT-B/16^{++}	70	0.811	0.831	**0.750**	0.790	0.910	0.594	0.752	0.869	0.663	0.877
	EfficientNetV2-S^{++}	45	0.768	0.787	0.706	0.746	0.919	0.436	0.678	0.846	0.512	0.811
	SwinV2-T^{++}	70	0.782	0.822	0.674	0.748	0.871	0.585	0.728	0.846	0.626	0.853
	MaxVit-T^{++}	33	0.801	0.836	0.715	0.776	0.886	0.615	0.751	0.860	0.658	0.867

Table 4. Results of Expt.#3. and Expt.#4. [1] Test Accuracy. [2] Benign class. [3] Malignant class. [4] Macro Avg. for Precision. [5] Macro Avg. for Recall

much longer time to complete. The best performance for the models is the InceptionV3, with overall accuracy equals 96.4% and AUC equals 99.3%. SwinV2-T is sensitive to the initial weights since it can be easily affected.

In conclusion, compared to training from scratch, the significant advantage of using ImageNet as the source domain is the reduction in running time, and indeed, other metrics are generally higher than training from scratch. This suggests that this target domain requires a larger and more diverse source domain to better capture features. However, whether it must specifically be ImageNet remains uncertain.

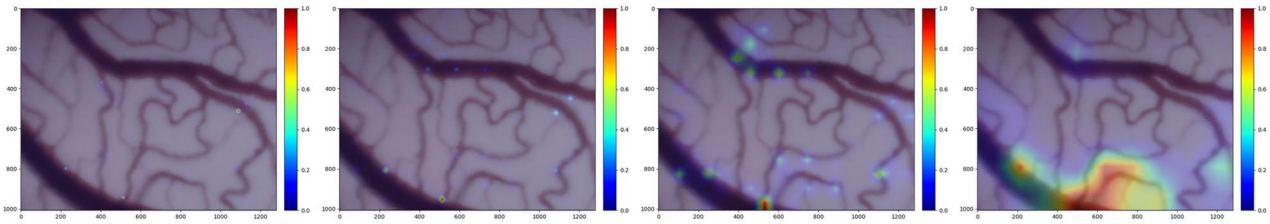
Results to verify Hypothesis (2)

The results of Expt.#3 and Expt.#4 are presented in Table 4. In the table, every record is the average of three runs, the number of epochs is rounded up, and the rest of the data is kept to three decimal places. For both experiments, the highest result is marked bold. The ROC curves for Expt.#3 and Expt.#4 are presented in the Figure 5c and Figure 5d. Given that we conduct three runs for each experiment, we select the run that is closest to the average testing accuracy presented in the Table 4 in order to generate the ROC curve.

Expt.#3 investigates the effectiveness of one-step heterogeneous transfer learning. Since only the fully-connected layer of the model is updated in the target domain, the model's ability to discriminate between target domain classes relies heavily on the knowledge learned from the source domain. MaxVit-T performs worst in classification tasks with an overall accuracy of 80.9%. On the other hand, ViT-B/16 performs best in this scenario with an overall accuracy of 91.5%, as well as the best performed classifier from the ROC curves in Figure 5c, although not as well in terms of efficiency.

Model	Epoch	Val. Acc.[1]	Avg. Time [2]	Model	Epoch	Val. Acc.	Avg. Time
DenseNet121^{+}	26	0.791	136s	ResNet50^{+}	38	0.743	106s
DenseNet169^{+}	26	0.777	156s	VGG19^{+}	32	0.738	150s
InceptionV3^{+}	20	0.822	142s	ViT-B/16^{+}	15	0.759	237s
MobileNet V2^{+}	17	0.808	120s	EfficientNetV2-S^{+}	13	0.84	251s
ResNet18^{+}	19	0.773	121s	SwinV2-T^{+}	16	0.825	543s
MaxVit-T^{+}	13	0.836	548s				

Table 5. Intermediate results of Model^{+} in Expt.#4 [1]Validation accuracy.[2] Average time per epoch, and is rounded up



(a) Layer 1 **(b)** Layer 2 **(c)** Layer 3 **(d)** Layer 4

Fig. 6. Visualize the area of interest of ResNet50** on benign example using LayerCAM.

Expt.#4 investigates the effectiveness of two-step heterogeneous transfer learning using the DR images as the intermediate domain. This intermediate domain is larger and more diverse than the target domain in terms of the number of images and classes. The intermediate results are summarized in Table 5.

VGG19, ViT-B/16, and SwinV2-T all reach the maximum epoch limit, but VGG19 outperforms overall, achieving an accuracy of 83.9% and emerging as the best classifier in Figure 5d. It exhibits the highest recall for malignant class, indicating its superior ability to accurately identify malignant cases. Combining the models' performances from Table 5 and Table 4, achieving the best performance in the intermediate domain does not guarantee the highest performance in the target domain. Additionally, it is important not to underestimate the precision rate of ViT-B/16 for the malignant class, as its high precision holds significant clinical value by reducing the likelihood of missing actual afflicted patients. Among the models, MobileNet V2 proves to be the most time-efficient; however, it performs the poorest in the laryngeal blood vessel classification task, with an accuracy of only 67.8%.

Overall, after learning in the intermediate domain, there is a general improvement in the model's benign recall of the target domain by about 1.4% on average, which indicates that the features of the intermediate domain contribute to identifying the benign class in the target domain. However, there is a substantial drop in the overall performance of THTL compared with the one-step in this classification task. The average accuracy of 11 models drops by 10.6%, and the average AUC drops around 13.6%.

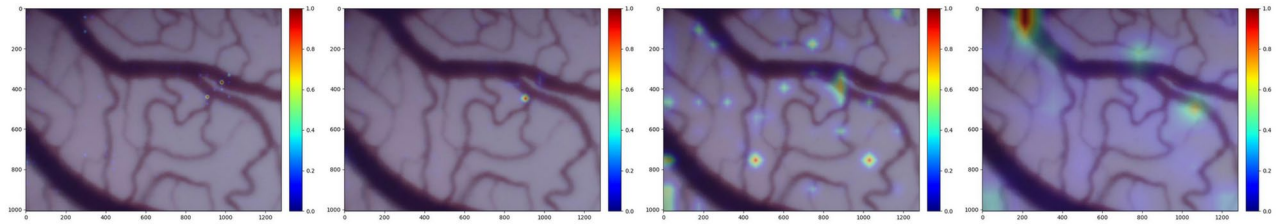
It can be observed that the overall performance declined in Expt.#4, with the malignant class contributing the most. The average percentage decrease in the malignant class is more significant than in the benign class for almost every metric, especially for precision and recall. This suggests that the same model pre-trained on ImageNet loses most of its capability to determine the malignant class after learning the vessel pattern from the intermediate domain. The possible reason is identified by the visualization of the model's attention and will be discussed in Analysis of Performance Drop using LayerCAM.

Generally speaking, all the models are more capable of identifying the benign class than the malignant class, regardless of the presence of an intermediate domain or whether they are trained from scratch. EfficientNetV2-S performs well when it requires well-prepared initial weights for its network and is thoroughly trained on the target domain. For training from scratch, InceptionV3 is more suitable in this scenario. Mainly, MaxVit-T shows excellent ability in one-step heterogeneous transfer learning. VGG19 is advanced in two-step heterogeneous transfer learning among the 11 models.

Analysis of performance drop using LayerCAM

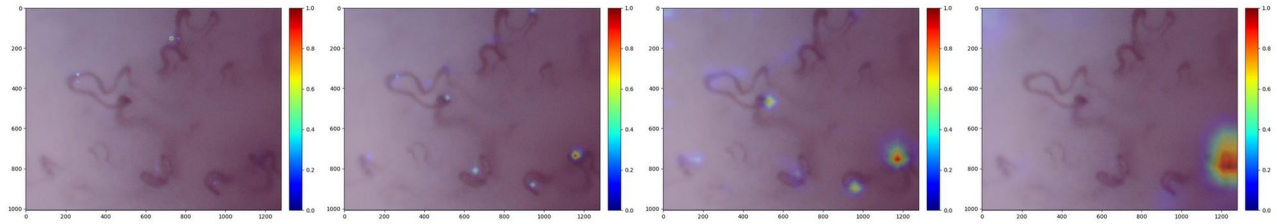
For a long time, the actual inner operation of deep learning models has been a black box. However, with the help of the advent of attention mechanisms and the development of model visualization, it has allowed us to finally glimpse the inner mysteries of the models.

After comprehensive analysis, we conclude a similar trend within these images. We randomly select a benign-labeled image (image number Patient012_P012 (33)) from the test set. Then the LayerCAM is used to visualize the benign class based on ResNet50** and ResNet50^{++}, shown in Figure 6 and Figure 7. A color bar ranging from 0 to 1 is added to the right side of each subfigure to indicate the level of attention the model assigns to each region of the image. Red represents regions where the model pays the highest attention, followed by yellow, while



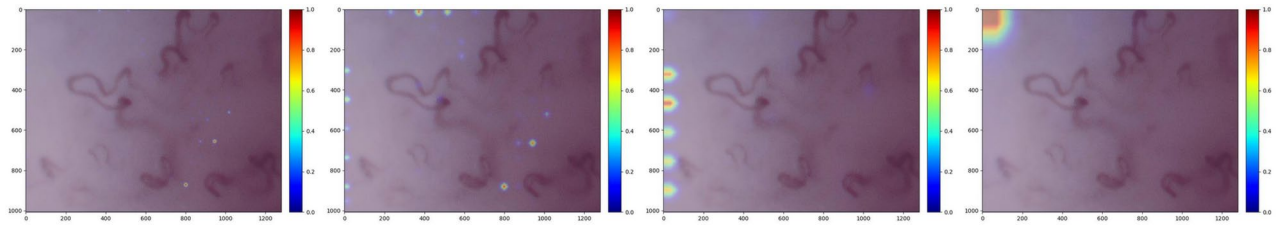
(a) Layer 1 **(b)** Layer 2 **(c)** Layer 3 **(d)** Layer 4

Fig. 7. Visualize the area of interest of ResNet50^{++} on benign example using LayerCAM.



(a) Layer 1 **(b)** Layer 2 **(c)** Layer 3 **(d)** Layer 4

Fig. 8. Visualize the area of interest of ResNet50** on malignant example using LayerCAM.



(a) Layer 1 **(b)** Layer 2 **(c)** Layer 3 **(d)** Layer 4

Fig. 9. Visualize the area of interest of ResNet50^{++} on malignant example using LayerCAM.

blue indicates regions with the least attention. Both ResNet50** and ResNet50^{++} deliver a correct prediction on this image. It can be observed that ResNet50** starts focusing more on the contours of the major vessels and gradually moving the focus to the lower corner of the image as the number of layers increases, with the thicker vessels being the primary basis for judgment. ResNet50^{++} applies the acquired vascular knowledge to the classification prediction task of the target domain by learning the intermediate domain, it puts more effort into capturing the fine vessels used as a basis for judgment. It can be inferred that after learning in the intermediate domain, the ResNet50 holds the ability to locate the fine vessels.

A malignant-labeled image is randomly selected from the test set, with image number Patient063_P063 (26). The LayerCAM visualization of the malignant class based on ResNet50** and ResNet50^{++} are represented in Figure 8 and Figure 9. Unlike the benign classification, this time, ResNet50** gives a correct prediction for the image while ResNet50^{++} does not. As can be seen in Figure 8a, ResNet50** puts its attention on the edges of the twisted blood vessels at the beginning. As the number of layers increases, the model concentrates more on the twisted or tadpole-shaped blood vessels which are highlighted by LayerCAM. However, for ResNet50^{++}, the model increasingly focuses on empty spaces within the image from Layer 1 to Layer 2 and shifts entirely to the edges by Layer 3. Finally, ResNet50^{++} focuses on the blank space at the upper left corner of the image. The comparison between Figure 8 and Figure 9 reveals that after learning in the intermediate domain, ResNet50^{++} fails to acquire the ability to distinguish twisted and entangled blood vessels. Moreover, by Layer 3, it completely loses the capacity to identify the malignant class. Therefore, the classification performance in this class drops.

To sum up, images from the target dataset do not always follow a radial pattern, and some images show a large angle bend for blood vessels. Through the analysis of LayerCAM visualization outcomes, we speculate that in two-step heterogeneous transfer learning, ResNet50 loses the ability to discriminate twisted and tangled vessels while gaining the ability to capture fine vessels. This speculation applies to generalize to other models that have significant performance drops for predicting the malignant class in Expt.#4.

Step	Epoch	Acc.[1]	Precision			Recall			F1-Score		AUC
			Benign	Malignant	MA[2]	Benign	Malignant	MA[3]	Benign	Malignant	
1	44	0.738	0.743	0.702	0.723	0.946	0.282	0.614	0.832	0.402	0.766
2	29	0.901	0.925	0.845	0.885	0.931	0.834	0.883	0.928	0.84	0.957
3	34	**0.942**	**0.957**	0.908	**0.933**	**0.959**	**0.904**	**0.931**	**0.958**	**0.906**	0.979
4	29	0.937	0.95	0.906	0.928	0.958	0.889	0.924	0.954	0.897	0.982
5	32	0.938	0.949	**0.913**	0.931	0.962	0.885	0.923	0.955	0.899	**0.987**
6	30	0.938	0.952	0.907	0.930	0.958	0.895	0.926	0.955	0.901	0.981
1 (LoRA)	40	0.736	0.748	0.670	0.709	0.931	0.309	0.620	0.829	0.423	0.771
2 (LoRA)	51	0.899	0.923	0.847	0.885	0.931	0.829	0.880	0.927	0.837	0.958
3 (LoRA)	41	0.921	0.941	0.877	0.909	0.944	0.870	0.907	0.943	0.873	0.974
4 (LoRA)	32	0.920	0.938	**0.880**	0.909	**0.946**	0.862	0.904	0.942	0.871	0.972
5 (LoRA)	41	**0.925**	0.946	**0.880**	**0.913**	0.945	0.881	**0.913**	**0.946**	**0.880**	**0.978**
6 (LoRA)	32	0.922	**0.949**	0.866	0.907	0.937	**0.889**	**0.913**	0.943	0.877	0.973

Table 6. Results of Step-Wise Fine-Tuning for ResNet50^{++} verses LoRA for ResNet50^{++} [1] Test Accuracy. [2] Macro Avg. for Precision. [3] Macro Avg. for Recall

Step-wise fine-tuning

Since we use LayerCAM to visualize the attention areas for ResNet50, to better cooperate with the LayerCAM visualization and ResNet's modules in Pytorch, we propose a fine-tuning technique called Step-Wise Fine-Tuning. SWFT compensates for the absence of features in the intermediate domains that can be supported to classify the malignant class on the target domains, and the accuracy even exceeds updating all layers in THTL when fine-tuning to Step 3.

Following this Step-Wise Fine-Tuning design, we conduct experiments with ResNet50^{++} on the target domain. The metrics from Expt.#4 are extended here, and similarly, the average value of three runs is recorded for each metric, then presented in Table 6.

It can be observed from Table 6 that the performance of the classification task improves significantly from Step 1 to Step 2, especially for malignant recall and F1-Score. From the metrics perspective, only fine-tuning the fully connected layer is less efficient. It takes the longest time to complete the task and performs least unsatisfied.

Specifically, our proposed SWFT also incorporates the visualization of the LayerCAM as a reference for the number of steps that should be involved in SWFT. As illustrated in Figure 9, after the intermediate domain, ResNet50^{++} focuses on empty regions rather than the blood vessels. Also, most of the model's attention in Figure 9c focuses on the edge of the image instead of vessels. This observation suggests that at Layer 3, the model is no longer focused on the vessel itself, indicating that the SWFT should be fine-tuned up to Step 3. The efficacy of this approach is substantiated by Table 6, which demonstrates that the accuracy achieved at Step 3 surpasses not only the accuracy of step 1 (increases by 20.4%) but even that of fine-tuning the entire network, as well as precision of benign and malignant, recall of benign and malignant, and F1-Score of benign and malignant. Still, this performance is 1.6% less accurate compared to Expt.#1, but it is close.

Furthermore, we compare our result to the adaptive fine-tuning method Low-Rank Adaptation (LoRA)[56] in the second step of THTL. LoRA is highly regarded in large language models for its efficiency and extremely small parameter overhead. First, we freeze the weights for the entire intermediate-domain-trained model, and use LoRA (with $rank = 8, \alpha = 16$, initial kaiming uniform and $a = \sqrt{5}$ for matrix A and initial zeros for matrix B) to fine-tune the whole ResNet50^{++} in the target domain ("Step 6 (LoRA)"in Table 6). This result is lower than the result obtained by our proposed method (see"Step 6"in the same table). To ensure a fairer comparison, we also integrate LoRA into our proposed fine-tuning method by progressively fine-tuning each Layer using LoRA from back to front, while keeping the ResNet50^{++} weights frozen. As seen in the results from Table 6, under the premise of fine-tuning with LoRA, fine-tuning the entire network does not lead to the best performance. The performance at "Step 5 (LoRA)"exceeds that of fine-tuning the entire network at"Step 6 (LoRA)".

All in all, the best performance of our proposed SWFT (94.2% accuracy) surpasses the best result from fine-tuning with LoRA (92.5% accuracy). Furthermore, fine-tuning the entire model dose not yield the best performance when LoRA is applied to the proposed SWFT. This further demonstrates that our proposed SWFT is practically feasible in THTL.

Conclusion

This work innovatively proposes two-step heterogeneous transfer learning for larynx cancer identification and validates using 11 deep learning models. Based on that, we summarise the key principles for intermediate domain selection for later studies. We use ImageNet as the source domain, the DR dataset as the intermediate domain, and the CE-NBI dataset as the target domain.

First, we verify hypothesis (1) and conclude that a larger and more diverse source domain would benefit larynx cancer identification. Based on this derivation and emphasizing the clear representation of blood vessel properties while maintaining heterogeneity, we selected diabetic retinopathy images as the intermediate domain for THTL. This dataset is larger in scale and has more classes than CE-NBI. However, hypothesis (2) gives

a contrasting result than expected. The experiment results demonstrate that models diminish the ability to distinguish the malignant class after learning in the intermediate domain, reflected in the evaluation metrics, such as precision, recall, AUC, and accuracy. With the help of LayerCAM, we visualize the decision-making process of the model layer by layer, finding that the model struggles to pay attention to the twisted, tangled vessels, which are the essential character for distinguishing the malignant class in the target domain. The visualization of LayerCAM also reflects different features of the blood vessels in two domains. Most of the vessels in the intermediate domains show a radial pattern, which is favorable for supporting identifying the benign class in the target domain but not the malignant class. Therefore, we summarize two principles for intermediate domain selection for using THTL in larynx cancer identification: 1) the intermediate domain does not necessarily have to be larger in scale or number of classes than the target domain, as the features are a more dominant factor; 2) radial vascular patterns support the classification of benign cases in the target domain, while twisted and tangled patterns are more related to the malignant classification. Furthermore, we propose a Step-Wise Fine-Tuning technique, guided by the visualization result of LayerCAM, which significantly improves the performance in the second step of THTL and even exceeds the performance of fine-tuning all the layers of the model.

Despite our efforts, some limitations still exist within our work. One of the biases affecting the results could be the uniform vascular features in the intermediate domain. In subsequent laryngeal vascular classification studies, the intermediate domain should contain more shapes of vascular features to better serve the target domain. Another possible bias is that we focus on supervised learning in this work, though it enables the model to learn more accurate knowledge; however, at the same time, it limits the data that we can use. Therefore, in the future, it is worth exploring other learning approaches, such as self-supervised or unsupervised learning, to investigate more possibilities.

Data Availability

The datasets analysed during the current study are public available in Zenodo repository at https://zenodo.org/records/6674034, reference number[43]; and in Kaggle repository at https://www.kaggle.com/c/diabetic-retinopathy-detection/data, reference number[12].

Received: 15 October 2024; Accepted: 27 December 2024
Published online: 16 January 2025

References

1. Deng, Y. et al. Global burden of larynx cancer, 1990-2017: estimates from the global burden of disease 2017 study. Aging (Albany NY) **12**, 2545 (2020).
2. Huang, P. et al. La-vit: A network with transformers constrained by learned-parameter-free attention for interpretable grading in a new laryngeal histopathology image dataset. IEEE Journal of Biomedical and Health Informatics (2024).
3. Huang, P. et al. Mamlformer: Priori-experience guiding transformer network via manifold adversarial multi-modal learning for laryngeal histopathological grading. *Information Fusion* **108**, 102333 (2024).
4. Huang, P. et al. A vit-amc network with adaptive model fusion and multiobjective optimization for interpretable laryngeal tumor grading from histopathological images. *IEEE Transactions on Medical Imaging* **42**, 15–28 (2022).
5. Lukes, P. et al. Narrow band imaging (nbi)-endoscopic method for detection of head and neck cancer. *Endoscopy* **5**, 75–87 (2013).
6. Chabrillac, E. et al. Narrow-band imaging in oncologic otorhinolaryngology: State of the art. *European Annals of Otorhinolaryngology, Head and Neck Diseases* **138**, 451–458 (2021).
7. He, Y. et al. A deep convolutional neural network-based method for laryngeal squamous cell carcinoma diagnosis. Annals of Translational Medicine **9** (2021).
8. Oquab, M., Bottou, L., Laptev, I. & Sivic, J. Learning and transferring mid-level image representations using convolutional neural networks. In 2014 IEEE Conference on Computer Vision and Pattern Recognition, 1717–1724, https://doi.org/10.1109/CVPR.2014.222 (2014).
9. Tan, B., Zhang, Y., Pan, S. & Yang, Q. Distant domain transfer learning. In Proceedings of the AAAI conference on artificial intelligence, vol. 31 (2017).
10. Matsoukas, C., Haslum, J. F., Sorkhei, M., Söderberg, M. & Smith, K. What makes transfer learning work for medical images: Feature reuse & other factors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 9225–9234 (2022).
11. Raghu, M., Zhang, C., Kleinberg, J. & Bengio, S. Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems **32** (2019).
12. Dugas, E., Jared, Jorge & Cukierski, W. Diabetic retinopathy detection (2015).
13. Jiang, P.-T., Zhang, C.-B., Hou, Q., Cheng, M.-M. & Wei, Y. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing* **30**, 5875–5888. https://doi.org/10.1109/TIP.2021.3089943 (2021).
14. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* **22**, 1345–1359 (2010).
15. Day, O. & Khoshgoftaar, T. M. A survey on heterogeneous transfer learning. *Journal of Big Data* **4**, 1–42 (2017).
16. Otović, E. et al. Intra-domain and cross-domain transfer learning for time series data-how transferable are the features?. *Knowledge-Based Systems* **239**, 107976 (2022).
17. Bukhsh, Z. A., Jansen, N. & Saeed, A. Damage detection using in-domain and cross-domain transfer learning. *Neural Computing and Applications* **33**, 16921–16936 (2021).
18. Araújo, T., Santos, C. P., De Momi, E. & Moccia, S. Learned and handcrafted features for early-stage laryngeal scc diagnosis. *Medical & Biological Engineering & Computing* **57**, 2683–2692 (2019).
19. Xiong, H. et al. Computer-aided diagnosis of laryngeal cancer via deep learning based on laryngoscopic images. *EBioMedicine* **48**, 92–99 (2019).
20. Heker, M. & Greenspan, H. Joint liver lesion segmentation and classification via transfer learning. arXiv preprint arXiv:2004.12352 (2020).
21. Alzubaidi, L. et al. Optimizing the performance of breast cancer classification by employing the same domain transfer learning from hybrid deep convolutional neural network model. *Electronics* **9**, 445 (2020).
22. Xie, Y. & Richmond, D. Pre-training on grayscale imagenet improves medical image classification. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops (2018).
23. Ruder, S. Neural transfer learning for natural language processing. Ph.D. thesis, NUI Galway (2019).

24. De Matos, J., Britto, A. d. S., Oliveira, L. E. & Koerich, A. L. Double transfer learning for breast cancer histopathologic image classification. In 2019 international joint conference on neural networks (IJCNN), 1–8 (IEEE, 2019).
25. Alkhaleefah, M. et al. Double-shot transfer learning for breast cancer classification from x-ray images. *Applied Sciences* **10**, 3999 (2020).
26. Alzubaidi, L. et al. Novel transfer learning approach for medical imaging with limited labeled data. *Cancers* **13**, 1590 (2021).
27. Meng, J., Tan, Z., Yu, Y., Wang, P. & Liu, S. Tl-med: A two-stage transfer learning recognition model for medical images of covid-19. *Biocybernetics and Biomedical Engineering* **42**, 842–855 (2022).
28. Morid, M. A., Borjali, A. & Del Fiol, G. A scoping review of transfer learning research on medical image analysis using imagenet. *Computers in biology and medicine* **128**, 104115 (2021).
29. Szegedy, C. et al. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015).
30. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, 770–778 (2016).
31. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).
32. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016).
33. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 4700–4708 (2017).
34. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. & Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018).
35. Arfan, T. H., Hayaty, M. & Hadinegoro, A. Classification of brain tumours types based on mri images using mobilenet. In 2021 2nd International Conference on Innovative and Creative Information Technology (ICITech), 69–73, https://doi.org/10.1109/ICITech50181.2021.9590183 (2021).
36. Cho, W. K. et al. Diagnostic accuracies of laryngeal diseases using a convolutional neural network-based image classification system. *The Laryngoscope* **131**, 2558–2566 (2021).
37. Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (2021).
38. Cao, R., Shang, H., Zhang, L., Wu, L. & Zheng, Y. Transformer for computer-aided diagnosis of laryngeal carcinoma in pcle images. In 2021 International Conference on Networking Systems of AI (INSAI), 181–188, https://doi.org/10.1109/INSAI54028.2021.00042 (2021).
39. Tan, M. & Le, Q. Efficientnetv2: Smaller models and faster training. In Meila, M. & Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, vol. 139 of Proceedings of Machine Learning Research, 10096–10106 (PMLR, 2021).
40. Liu, Z. et al. Swin transformer v2: Scaling up capacity and resolution. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 12009–12019 (2022).
41. Tu, Z. et al. Maxvit: Multi-axis vision transformer. In European conference on computer vision, 459–479 (Springer, 2022).
42. Deng, J. et al. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–255, https://doi.org/10.1109/CVPR.2009.5206848 (2009).
43. Esmaeili, N. et al. Contact Endoscopy-Narrow Band Imaging (CE-NBI) Data Set for Laryngeal Lesion Assessment, https://doi.org/10.5281/zenodo.6674034 (2022).
44. Paszke, A. et al. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H. et al. (eds.) Advances in Neural Information Processing Systems, vol. 32 (Curran Associates, Inc., 2019).
45. Kingma, D. P. & Ba, J (A method for stochastic optimization, Adam, 2017) (**1412.6980**).
46. Afify, H. M., Mohammed, K. K. & Hassanien, A. E. Novel prediction model on oscc histopathological images via deep transfer learning combined with grad-cam interpretation. *Biomedical Signal Processing and Control* **83**, 104704 (2023).
47. Wang, Y. et al. A gradient mapping guided explainable deep neural network for extracapsular extension identification in 3d head and neck cancer computed tomography images. *Medical Physics* **51**, 2007–2019 (2024).
48. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2921–2929 (2016).
49. Kandel, I. & Castelli, M. Transfer learning with convolutional neural networks for diabetic retinopathy image classification. a review. Applied Sciences **10**, 2021 (2020).
50. Vrbančič, G. & Podgorelec, V. Transfer learning with adaptive fine-tuning. *IEEE Access* **8**, 196197–196211. https://doi.org/10.1109/ACCESS.2020.3034343 (2020).
51. Boumaraf, S., Liu, X., Zheng, Z., Ma, X. & Ferkous, C. A new transfer learning based approach to magnification dependent and independent classification of breast cancer in histopathological images. *Biomedical Signal Processing and Control* **63**, 102192 (2021).
52. Tajbakhsh, N. et al. Convolutional neural networks for medical image analysis: Full training or fine tuning?. *IEEE transactions on medical imaging* **35**, 1299–1312 (2016).
53. Sharma, S. & Mehra, R. Effect of layer-wise fine-tuning in magnification-dependent classification of breast cancer histopathological image. *The Visual Computer* **36**, 1755–1769 (2020).
54. Zeiler M. D. & Fergus R. Visualizing and understanding convolutional networks. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, 818–833 (Springer, 2014).
55. Ma C., Huang J.-B, Yang X. & Yang, M.-H. Hierarchical convolutional features for visual tracking. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015).
56. Hu E. J. et al. LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations (2022).

Acknowledgements

We would like to express our gratitude to Doctor Zhang Binghuang, the attending physician of the Department of Otolaryngology Head and Neck Surgery from the First Affiliated Hospital of Xiamen University, for his professional advice.

Author contributions

X.F. contributed to the concept, experiments, data analysis, and manuscript. C.F.C. and K.L.W. built and validated the model. M.S. contributed to data analysis and revised the manuscript. B.K.N. revised the manuscript and provided resources. All authors reviewed the manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to B.K.N.

Reprints and permissions information is available at www.nature.com/reprints.

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