Embracing Large Natural Data: Enhancing Medical Image Analysis via Cross-domain Fine-tuning

Qiankun Li, Xiaolong Huang, Bo Fang, Huabao Chen, Siyuan Ding, Xu Liu

***Abstract*—With the rapid advancements of big data and computer vision, many large-scale natural visual datasets are proposed, such as ImageNet-21K, LAION-400M, and LAION-2B. These large-scale datasets significantly improve the robustness and accuracy of models in the natural vision domain. However, the field of medical images continues to face limitations due to relatively small-scale datasets. In this paper, we propose a novel method to enhance medical image analysis across domains by leveraging pre-trained models on large natural datasets. Specifically, a Cross-Domain Transfer Module (CDTM) is proposed to transfer natural vision domain features to the medical image domain, facilitating efficient fine-tuning of models pre-trained on large datasets. In addition, we design a Staged Fine-Tuning (SFT) strategy in conjunction with CDTM to further improve the model performance. Experimental results demonstrate that our method achieves state-of-the-art performance on multiple medical image datasets through efficient fine-tuning of models pre-trained on large natural datasets.**

***Index Terms*—Large natural data, medical image, Cross-domain learning, Staged fine-tuning**

I. Introduction

With the rapid advancement of artificial intelligence, the field of computer vision has witnessed the emergence of numerous advanced algorithms based on deep learning [1], [2]. These algorithms are data-driven, meaning they leverage large volumes of data to enhance their performance on specific tasks [3], [4]. The accumulation of information in the digital age has recently led to the creation of many large-scale natural datasets [5]–[7], designed to improve the overall accuracy and versatility of algorithms by providing diverse and extensive data. These large datasets have propelled numerous industrial applications in natural scenarios, such as service robots [8], autonomous driving [9], and intelligent security systems [10]. However, the medical imaging domain still faces the challenge of limited data, impeding the progress and widespread application of artificial intelligence in the medical field compared to natural scenarios.

To enhance the performance of deep learning algorithms on limited medical images, some studies have employed transfer learning techniques, utilizing the pre-training-then-fine-tuning paradigm. Raghu *et al.* [11] conducted pre-training on approximately 5,000 medical chest X-Rays [12] and 9,000 retinal fundus [13] images, then performed fine-tuning on the same domain dataset. Alzubaidi *et al.* [14] proposed an unsupervised learning approach, wherein they pre-trained the model on 200,000 unlabeled skin cancer images [15]–[17] and subsequently fine-tuned it on a small labeled skin cancer dataset. However, even with the inclusion of unlabeled images, the scale of these medical datasets still falls short compared to natural datasets. He *et al.* [18], Liu *et al.* [19], and McKinney *et al.* [20]have resorted to pre-training models on the extensively utilized ImageNet-1K dataset, which encompasses around 1.3 million natural images. Subsequently, they fine-tuned models for downstream medical tasks. However, Neyshabur *et al.* [21] and Heker *et al.* [22] argue that there are significant domain differences between medical and natural images, which can limit the performance of cross-domain fine-tuning.

In this paper, we propose a novel method to transfer rich features from large-scale natural vision datasets to medical images (As shown in Fig. 1). Concerning pre-training large natural datasets, the study spans the main dimensions of modern. This includes the ten-million level ImageNet-21K [6]

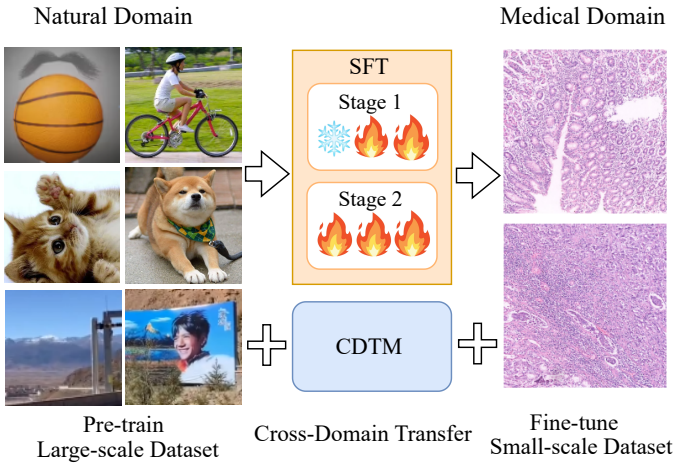

Fig. 1. Overview of the proposed method. It aims to transfer a pre-trained model with rich features from large-scale natural datasets to data-limited medical image datasets. A Cross-Domain Transfer Module is proposed to address the significant difference between the two domains. The Staged Fine-Tuning (SFT) strategy is designed to collaborate with CDTM and improve the fine-tuning process. In Stage 1, the pre-trained backbone is frozen while CDTM and head undergo fine-tuning. In Stage 2, the backbone is unfrozen for full fine-tuning.

Qiankun Li is with Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, China, and also with the Department of Automation, University of Science and Technology of China, Hefei 230027, China.

Xiaolong Huang is with School of Artificial Intelligence, Chongqing University of Technology, Chongqing 401120, China.

Bo Fang is with College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110004, China.

Huabao Chen is with College of Energy and Electrical Engineering, Hohai University, Nanjing, 211100, China.

Siyuan Ding and Xu Liu are both with Department of Gastroenterology, General Hospital of Northern Theatre Command, 83 Wenhua Road, Shenyang, Liaoning 110840, China.

Xu Liu is the corresponding author (liuxu.north@outlook.com).

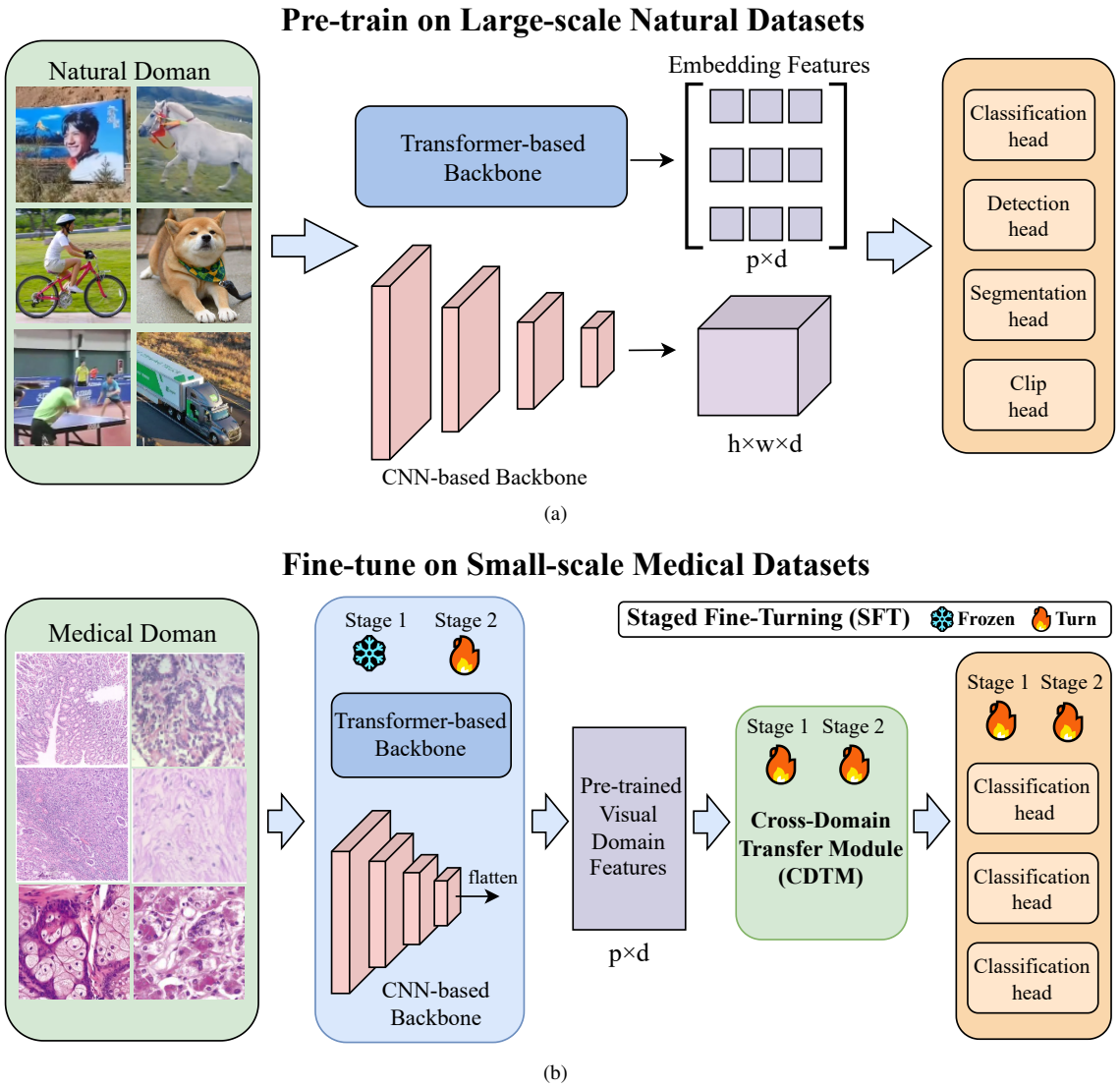

Fig. 2. Overview of the proposed method: (a) Pre-training on large-scale natural datasets; (b) Fine-tuning on small-scale medical datasets. The proposed method strives to enhance medical image analysis via cross-domain fine-tuning. This is accomplished by incorporating the Cross-Domain Transfer Module (CDTM) and the Staged Fine-Tuning (SFT) strategy.

classification dataset, the hundred-million level LAION-400M [7] image-text dataset, and the billion-level LAION-2B [5] image-text dataset. To address the challenging difficulties of cross-domain transfer learning, we propose a Cross-Domain Transfer Module (CDTM). This module introduces a cluster center attention mechanism to integrate and filter the output feature vectors of the pre-trained model, thereby facilitating cross-domain fine-tuning. Furthermore, a Staged Fine-Tuning (SFT) strategy is designed to collaborate with CDTM to further enhance the fine-tuning performance of the model on the medical dataset. In the first stage of SFT, the pre-trained backbone is frozen, and CDTM and head are fine-tuned to obtain better pre-transfer weights. In the second stage, the backbone is unfrozen for full fine-tuning. Our method achieves state-of-the-art performance on two popular medical image benchmarks, namely BreakHis [23] and HCRF [24], with F1 scores of 91.91% and 99.29%, respectively. Validation experiments and ablation analysis demonstrate the advantages of large-scale natural vision datasets and the effectiveness of cross-domain fine-tuning.

The contributions of this paper are summarized as follows.

1) We propose a novel method that utilizes pre-training on large-scale natural datasets to enhance medical image analysis through cross-domain fine-tuning.
2) A Cross-Domain Transfer Module (CDTM) is presented to address the significant differences between the natural images domain and the medical images domain.
3) A Staged Fine-Tuning (SFT) strategy is designed to collaborate with CDTM to further improve the fine-tuning performance of the model on downstream medical tasks.

The remainder of this paper is organized as follows. Section II details the architectural construction of the proposed method. The experiment configuration is introduced in Section III. Section IV describes the results and discussion, and the effectiveness of the propsed method is validated. Finally, Section V concludes this paper.

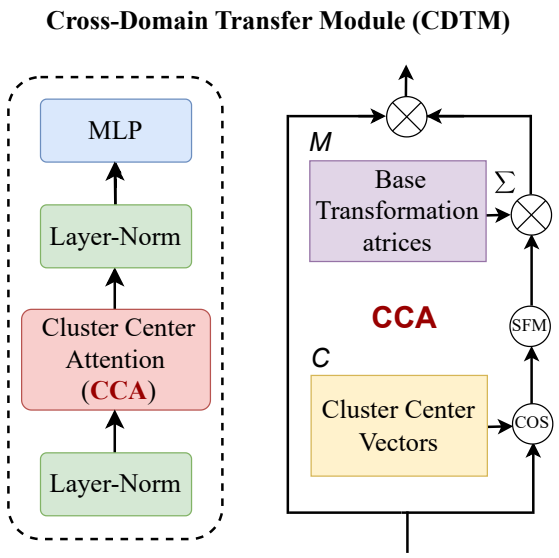

Fig. 3. The architecture of the Cross-DomainTransfer Module (CDTM). Cluster Center Attention (CCA) is the core component of CDTM.

II. Method

Fig. 2 is an illustration of the proposed method. It aims to transfer a pre-trained model with rich features from large-scale natural datasets to data-limited medical image datasets. This is accomplished by incorporating the Cross-Domain Transfer Module (CDTM) and the Staged Fine-Tuning (SFT) strategy.

A. Cross-Domain Transfer Module (CDTM)

The pre-trained models on large datasets in the natural visual domain possess rich features [25]–[27], and fine-tuning on downstream datasets yields impressive results [28], [29]. However, the current studies focus on image representations in the natural visual domain. To transfer the rich features from large natural data to medical images, it is crucial to address the significant differences between these two domains. In this direction, we propose a novel Cross-Domain Transfer Module (CDTM). This module introduces extra learnable parameters to the output features of a pre-trained model. It is based on the cluster center attention mechanism to integrate and filter the output feature vectors of the pre-trained model, preventing issues like gradient disappearance or explosion [30] that can arise from large-scale pre-trained features. Therefore, this module facilitates cross-domain fine-tuning more effectively.

1) Architecture: The architecture of the Cross-Domain Transfer Module (CDTM) is illustrated in Fig 3, which comprises Cluster Center Attention (CCA), Layer Normalization (Layer-Norm), and Multi-layer Perceptron (MLP). Pre-training models can be classified into two types, i.e., Transformer-based and CNN-based. The output size of Transformer-based models is $p \times d$, where p is the number of patches and d is the dimension of embedded features. Whereas CNN-based models have an output size of $h \times w \times d$, where h and w represent the width and height of the feature map, respectively, and d represents the dimension of embedded features. To unify the output size of both models, we can flatten the two-dimensional $h \times w$ of CNN-based models into one dimension, denoted as p as the same Transformer-based. This enables CDTM to perform cross-domain feature conversion uniformly on two types of pre-training models. The output of the pre-training model is first normalized by a Layer-Norm in CDTM, which reduces the difference in input distribution between different layers. This results in the pre-trained visual domain features x, with a size of $p \times d$. The core component CCA of CDTM is then used to complete the domain transformation. Finally, generalization and nonlinearity are enhanced through Layer-Norm [31] and MLP layers.

2) Cluster Center Attention (CCA): CCA is the core component of CDTM, as shown in Fig 3. CCA contains a set of learnable cluster center vectors C and a corresponding group of learnable base transformation matrices M, which can be defined as follows

$$C = \{c_i\}_{i=1}^{n}, \tag{1}$$

$$M = \{m_i\}_{i=1}^{n}, \tag{2}$$

where the variable n represents the number of cluster centers, c_i is the i-th cluster center vector of size $1 \times d$, and m_i is the i-th base transformation matrix of size $p \times p$. The variables d and p indicate the feature embedding dimension and patch numbers of the pre-trained domain features x, respectively.

The objective of CCA is to obtain the final cross-domain transformation matrix m_{cd} based on the features vector x, set of cluster center vectors C, and base transformation matrix group M. Specifically, CCA first calculates the global feature x_g of the feature vector x along the patch sequence dimension p, by

$$x_g = \sum_{j=1}^{p} x_j \tag{3}$$

where the variable x_j represents the embedding feature of j-th patch in features vector x. Then computes the cosine similarity between the global feature x_g and cluster center vectors C, and combines it with a scaling factor $\sqrt{d}$ to obtain the absolute attention score S_a. This process can be denoted as

$$s_{a_i} = \frac{x_g \cdot c_i}{\|x_g\| \, \|c_i\|} \quad (i = 1, 2, \ldots, n), \tag{4}$$

$$S_a = \frac{1}{\sqrt{d}} \cdot \{s_{a_i}\}_{i=1}^{n}, \tag{5}$$

where $\cdot$ denotes the dot product between vectors, and $\|\cdot\|$ denotes the norm of a vector. Taking into account the distance metric scale between the global feature x_g and all cluster centers C, the Softmax function [32] converts absolute attention scores S_a into relative attention scores S_r. This can be expressed as

$$S_r = softmax(S_a). \tag{6}$$

To achieve a cross-domain nonlinear transformation while maintaining the computational efficiency of the linear transformation, we apply weights sum to the base transformation matrices group M based on attention score S_r. This obtains

the cross-domain transformation matrix m_{cd}, defined as

$$m_{cd} = \sum_{i=1}^{n} s_{r_i} m_i. \tag{7}$$

Finally, the pre-trained natural visual domain feature x is transformed into a downstream medical image domain by the cross-domain transformation matrix m_{cd}, and enhanced with a Layer-Norm [31] and MLP layer to improve generalization and nonlinearity. The operation is mathematically denoted as follows

$$x_{cd} = MLP(LN(x \cdot m_{cd})), \tag{8}$$

where, x_{cd} is the output of the CDTM, which represents the features obtained after cross-domain transfer from the pre-trained natural visual domain to the medical image domain. The MLP layer comprises two fully connected layers separated by a GELU activation [33]. The first fully connected layer increases the feature dimension by four times, while the second fully connected layer reduces it back to its original dimension.

B. Fine-tuning Paradigm

After cross-domain transfer, the features obtained from the pre-trained natural visual model can be fed into the downstream medical image task head for fine-tuning. The commonly used fine-tuning methods are Linear Probing (LP) and Full Fine-tuning (FFT). Since the proposed CDTM introduces learnable cluster center vector groups C and basis transformation matrix M, we design a Staged Fine-tuning (SFT) strategy to better correspond CDTM. This section will explain each fine-tuning method in detail.

1) Linear Probing (LP): The LP method involves freezing the weights of the pre-trained backbone model while only training and updating the weights of the added CDTM and the head. Therefore, LP represents a trade-off between fine-tuning performance and computational efficiency. Taking the ViT-base network as an illustration, only approximately 0.002% of the overall network parameters are required for training. To better demonstrate the rich features provided by the pre-trained large data and the role of CDTM for domain transfer features, we use the simple linear layer for the head of medical image task.

2) Full Fine-tuning (FFT): The FFT method refers to the process of training and updating all parameters of the pre-trained model. In contrast to LP, FFT sacrifices some training efficiency to enhance model accuracy. However, the CDTM and linear head are trained from scratch, while the backbone of the model starts from pre-training weights. This simultaneous fine-tuning procedure can impede transfer learning of the CDTM.

3) Staged Fine-tuning (SFT): To address the problem between the FFT and CDTM, we design an SFT strategy. SFT combines the training process of LP and FFT. In the first stage, SFT freezes the weights of the pre-trained backbone while fine-tuning the CDTM and linear head. Subsequently, in the second stage, SFT unfreezes the backbone and fine-tunes the full model. This approach allows CDTM first to obtain better pre-transfer capabilities for the original pre-training domain, then further fine-tune the full model to complete the medical image task. Despite SFT not exhibiting the same training efficiency level as LP, it can optimize the transfer of rich features from the pre-training large dataset to the medical image to the greatest extent possible, thereby improving accuracy.

Table I. Details of the pre-training large natural datasets.

Dataset	Scale	Type	Task
ImageNet-21K	14 Million	images	classification
LAION-400M	400 Million	image-text pairs	multi-modal
LAION-2B	2000 Million	image-text pairs	multi-modal

III. Experiment Configuration

A. Pre-training Large Natural Datasets

The currently public natural visual large datasets consist of image classification and multi-modal image-text datasets. Taking into account the available pre-training models on these datasets, experiments select the ImageNet-21K classification dataset [34] at the ten-million level, the LAION-400M image-text dataset [7] at the hundred-million level, and the LAION-2B image-text dataset [5] at the billion level. Details of these datasets are listed in Table I.

1) ImageNet-21K: The ImageNet-21K [34] is a large-scale dataset with over 14 million natural images, covering over 21,000 categories. It was created by researchers at Princeton University and Stanford University, and is one of the largest and most commonly used datasets for image classification tasks in early computer vision. The ImageNet-21K dataset has been instrumental in advancing the state-of-the-art in computer vision and machine learning research.

2) LAION-400M: LAION-400M [7] is a dataset comprising 400 million pairs of CLIP image embeddings and their corresponding texts. The LAION team recently created this dataset and made it publicly available, which has significantly advanced the field of computer vision, particularly in large-scale pre-training, image-text matching, image generation, and text generation. In addition, the LAION-400M community also provides numerous available pre-training models that achieve impressive results in downstream tasks, including fine-tuning classification and even zero-shot learning.

3) LAION-2B: LAION-2B and LAION-400M are datasets based on the same benchmark, but the former is larger in scale, containing over 2 billion image-text pairs. However, the data quality of LAION-400M is improved due to the filtering and cleaning process it underwent with CLIP. Moreover, both datasets offer a similarly extensive range of pre-trained models.

B. Fine-tuning Medical Image Datasets

1) BreakHis: The BreakHis dataset [23] comprises 7,909 images at four different magnification levels and is divided into eight sub-classes of breast cancers. The source data was obtained from 82 anonymous patients at the Pathological

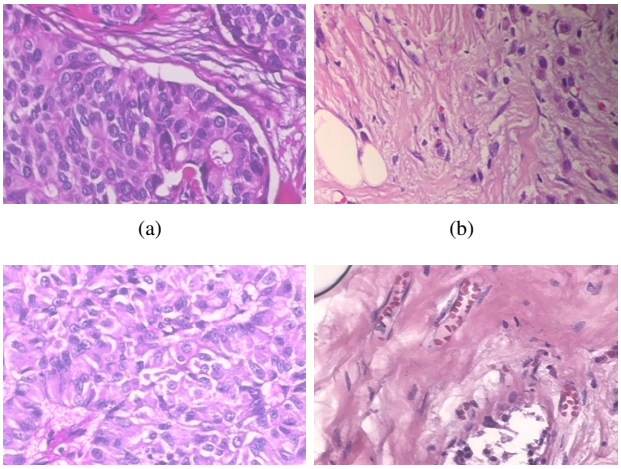

Fig. 4. An example of 200x malignant tumors in the BreakHis dataset: (a) ductal carcinoma (DC); lobular carcinoma (LC); (c) mucinous carcinoma (MC); (d) papillary carcinoma (PC).

Table II. Data setting of the BreakHis dataset.

Category	Sum	Training	Validation	Test
DC	896	538	179	179
LC	163	97	33	33
MC	196	118	39	39
PC	135	81	27	27

Anatomy and Cytopathology Lab in Brazil1. BreakHis is a well-known public dataset in the field of digital breast histopathology. It has been widely used in the development and evaluation of medical image analysis systems for breast cancer diagnosis. Following prior work [35], challenging malignant tumors with a magnification of $200\times$ are used for the four classifications, including ductal carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC) of the breast. An example of $200\times$ malignant tumors is shown in Fig. 4. As listed in Table II, the dataset is divided into training, validation, and testing sets in a ratio of 3:1:1 by random stratified.

2) HCRF: The HCRF dataset [24] is a popular gastric slice dataset with 560 cancerous images and 140 normal images. Some samples of normal and abnormal gastric histopathological images are shown in Fig. 5. As listed in Table III, the dataset is divided into training, validation, and test sets at a ratio of 1:1:2 by random stratified. The dataset is available on Mendeley Data and is suitable for evaluating model performance in the fields of computer vision and bioengineering.

Table III. Data setting of the HCRF dataset.

Category	Sum	Training	Validation	Test
Normal	140	35	35	70
Abnormal	140	35	35	70

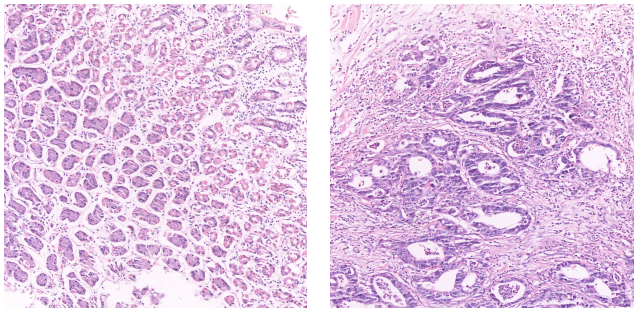

Fig. 5. Normal and cancerous gastric slice samples in the HCRF dataset: (a) normal sample; (b) cancerous sample.

C. Network Selection

To facilitate a more widespread transfer of large datasets from the natural vision field to the medical image field, we use the ViT [36] and ConvNeXt [37] pre-training models as experimental benchmarks, commonly provided in most large datasets. In addition, considering the equipment limitations of medical imaging systems in clinical applications, the specifications of the networks are chosen base size.

D. Evaluation Metric

To evaluate the performance of the proposed method, we utilize the precision, recall, accuracy, and F1 score as metrics. The calculation formulas of each metric are as follows

$$Pre = \frac{TP}{TP + FP}, \tag{9}$$

$$Rec = \frac{TP}{TP + FN}, \tag{10}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \tag{11}$$

$$F1 = \frac{2 \cdot Pre \cdot Rec}{Pre + Rec}, \tag{12}$$

where TP (True Positive) refers to the number of correctly predicted positive instances, TN (False Negative) represents the number of correctly predicted negative instances, FP (False Positive) indicates the number of actual negative instances that were incorrectly predicted as positive, and FN (True Negative) refers to the number of actual positive instances that were incorrectly predicted as negative.

E. Implementation Details

The input image size is set at 224×224, the learning rate is 1e-4. The maximum number of training epochs is 75, with a batch size of 16. The AdamW optimizer with a momentum of 0.9 and weight decay of 1e-3 is utilized for networks. The parameter n for cluster center in Eq. (1) of the Cross-Domain Transfer Module (CDTM) is set to 300. In our experimental results, we employ the efficiency-driven Linear Probing (LP) and accuracy-driven Staged Fine-tuning (SFT) paradigms for

Table IV. The experiment results on the BreakHis.

Method	Pre (%)	Rec (%)	Acc (%)	F1 (%)
Xception [38]	79.24	78.62	85.33	78.84
ResNet-50 [39]	74.60	76.88	82.66	75.54
Inception-V3 [40]	79.18	79.84	84.62	79.02
VGG-16 [41]	81.32	79.20	85.74	79.86
VGG-19 [41]	78.42	77.96	84.39	77.14
BotNet-50 [42]	79.20	80.72	85.32	79.50
GasHis-Transformer [35]	83.92	83.16	88.10	83.48
LW-GasHis-Transformer [35]	84.54	82.99	87.93	83.69
Transfer from LAION-2B				
ConvNeXt-B-CDTM-LP	80.78	75.09	84.17	77.83
ConvNeXt-B-CDTM-SFT	**93.09**	87.14	92.45	90.02
ViT-B-CDTM-LP	84.31	78.33	**87.41**	**81.21**
ViT-B-CDTM-SFT	92.74	**91.08**	**94.24**	**91.91**

fine-tuning. In the ablation experiments, we compare all fine-tuning methods together.

The experiments are performed on an Nvidia GeForce RTX 3090 GPU with 24 GB memory and Ubuntu 20.04 operating system. Python 3.8.3 serves as the programming language, while PyTorch 1.13.1 framework is employed. In addition, the source code is openly available on GitHub for interested readers.

IV. Results and Discussions

A. Results on BreakHis

Table IV lists the classification performance of fine-tuning pre-trained models transfer learning from the large natural visual LAION-2B dataset [5] on the downstream medical image BreakHis dataset [23]. It can be observed that the ViT-B-CDTM-SFT model (ViT Base model with our Cross-Domain Transfer Modul and Staged Fine-tuning) achieves a new state-of-the-art on the test set with average scores of 92.74%, 91.08%, 92.24%, and 91.91% for precision, recall, accuracy, and F1, respectively. In particular, the average F1 score of ViT-B-CDTM-SFT is 8.22% higher than the previously optimal LW-GasHis-Transformer, greatly enhancing the reliability of computer-aided breast cancer diagnosis. In addition, even the ViT-CDTM-LP trained with an efficiency-driven linear probing (LP) fine-tuning strategy achieves comparable results to the previously best-performing methods when only few of the parameters are trained with a frozen backbone. For the ConvNeXt-B-CDTM-SFT model, which also performs cross-domain transfer from the LAION-2B dataset, the F1 score is slightly lower than that of ViT-CDTM-SFT by 1.89% but still significantly outperforms previous methods.

In the confusion matrix shown in Fig. 6, the SFT strategy significantly improves the classification performance of the model for all types of cancer. Specifically, ConvNeXt-B-CDTM-SFT exhibits a remarkable increase in the classification accuracy of papillary carcinoma (PC) from 81% to 100%. Furthermore, ViT-B-CDTM-SFT enhances the recognition rate of challenging lobular carcinoma (LC) from 75% to 88%. In addition, our proposed methodology achieves the highest classification accuracy of 96% and 95% for ductal carcinoma (DC) and mucinous carcinoma (MC), respectively.

Table V. The experiment results on the HCRF.

Method	Pre (%)	Rec (%)	Acc (%)	F1 (%)
Xception [38]	94.48	97.78	95.94	95.98
ResNet-50 [39]	93.40	95.26	94.24	94.26
Inception-V3 [40]	93.64	94.40	93.80	93.96
VGG-16 [41]	90.82	94.48	92.34	92.38
VGG-19 [41]	88.68	94.68	91.24	91.34
BotNet-50 [42]	87.72	90.56	88.88	88.84
TransMed [43]	94.34	97.06	95.58	95.58
LeViT [44]	91.90	90.50	91.26	91.26
HCRF-AM [45]	92.90	91.94	94.24	92.06
SENetCNN [46]	95.94	95.94	95.94	95.94
CBAM-ResNet [47]	94.22	96.10	95.04	94.00
GasHis-Transformer [35]	98.55	97.38	97.97	97.97
LW-GasHis-Transformer [35]	95.99	96.90	96.43	96.43
Transfer from LAION-2B				
ConvNeXt-B-CDTM-LP	92.89	92.86	92.86	92.88
ConvNeXt-B-CDTM-SFT	**99.30**	**99.29**	**99.29**	**99.29**
ViT-B-CDTM-LP	87.17	87.14	87.14	87.16
ViT-B-CDTM-SFT	94.32	94.29	94.29	94.30

B. Results on HCRF

The HCRF dataset aims to identify gastric cancer and normal cases, with lower difficulty compared to the four-class cancer recognition in the BreakHis dataset. The classification results are listed in Table V. In comparison to numerous existing methods on this dataset, the proposed ConvNeXt-B-CDTM-SFT achieves a new state-of-the-art performance on the test set, with average precision, recall, accuracy, and F1 score of 99.30%, 99.29%, 99.29%, and 99.29% respectively. Notably, although the previously optimal GasHis-Transformer achieves a high F1 score of 97.97%, our model surpasses it by 1.32%, thus further enhancing the reliability of automatic gastric cancer diagnosis. In addition, ConvNeXt-B-CDTM-LP, which has minimal training parameters, achieves an F1 score of 92.88%. Furthermore, due to the limited size of downstream datasets, methods based on ViT-B perform less effectively than those based on ConvNeXt. However, ViT-B-CDTM-SFT still achieves comparable results to the previously best-performing methods.

C. Discussion on pre-training visual data of different scales

To further investigate the impact of the scale of large natural visual datasets on cross-domain fine-tuning, we conduct comparative experiments on the BreakHis dataset, using the ConvNeXt-B and ViT-B pre-trained models from three large-scale datasets, i.e., ImageNet-21K, LAION-400M, and LAION-2B. These models all incorporate CDTM and are fine-tuned using efficiency-driven LP and accuracy-driven SFT fine-tuning paradigms. The results are listed in Table VI

For the LP paradigm, the models under the largest LAION-2B dataset do not achieve the best results on medical images. Specifically, the ConvNeXt under ImageNet-21K and LAION-400M achieve comparable results, with F1-score higher than ConvNeXt under LAION-2B by 3.14% and 2.79%, respectively. This demonstrates that the quality of a large dataset is more important than its quantity when fine-tuning a few

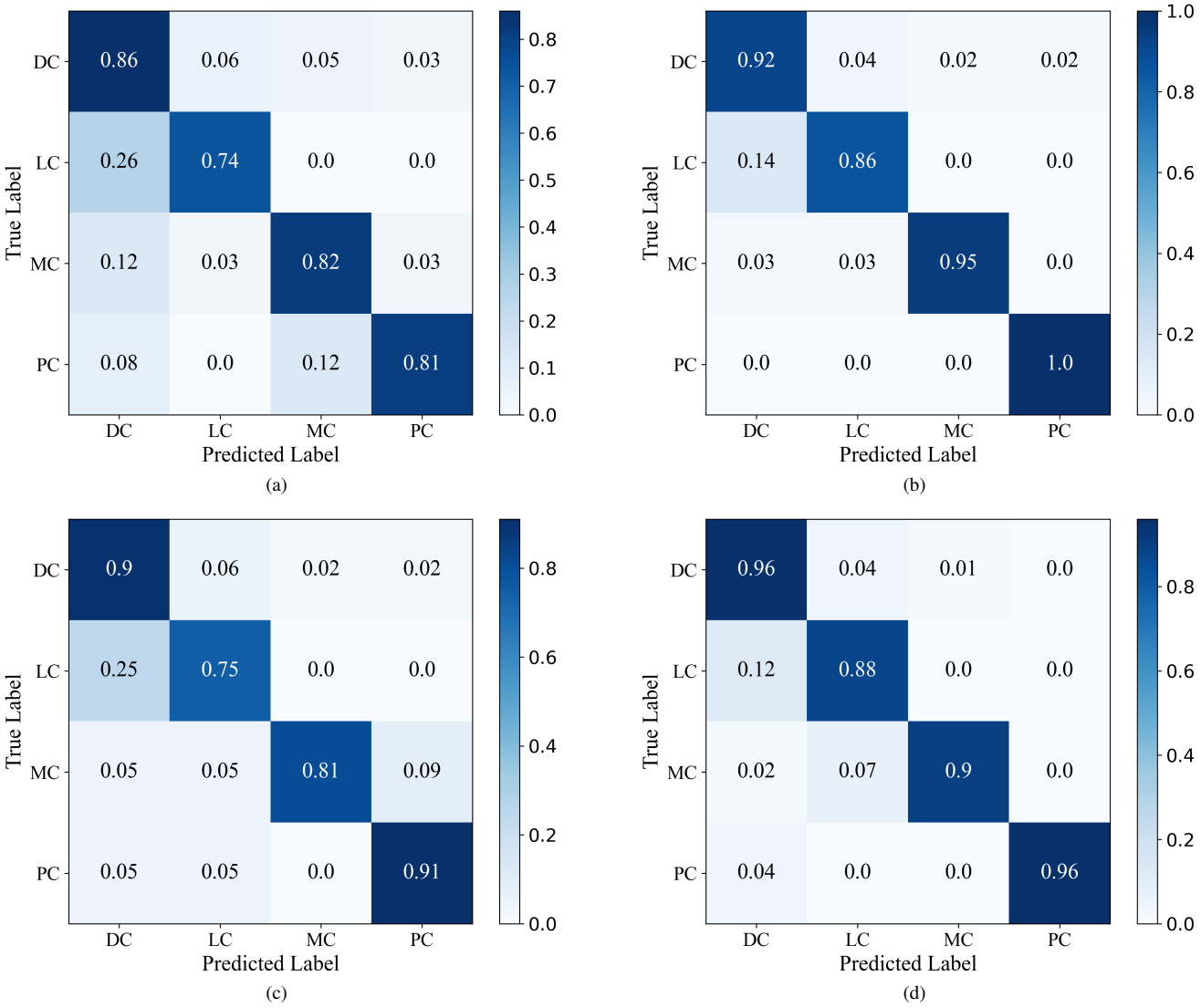

Fig. 6. Confusion matrix for different methods on BreakHis dataset: (a) ConvNeXt-B-CDTM-LP; (b) ConvNeXt-B-CDTM-SFT 2; (c) ViT-B-CDTM-LP; (d) ViT-B-CDTM-SFT.

parameters with a frozen backbone. This finding is further supported by the ViT model, as the F1 score of ViT-B under LAION-400M is 1.26% higher than that of ViT-B under LAION-2B. In addition, the performance of ViT-B under ImageNet-21K did not exceed that of ViT-B under the LAION dataset again, which is consistent with existing research [25], [48] that suggests ViT is skilled at handling multimodal datasets, such as the LAION series.

The importance of scale in the SFT paradigm becomes apparent when working with large datasets. It is noteworthy that both ConvNeXt and ViT models achieve the best cross-domain fine-tuning results on medical images when pre-trained on the largest-scale LAION-2B dataset. Even though the LAION-400M dataset has higher quality than LAION-2B, the F1 score of ConvNeXt on the medical dataset is still 1.6% lower than that of ConvNeXt under LAION-2B. In addition, the medical dataset F1 score of ViT under two pre-training LAION datasets is mostly the same, but ViT under LAION-2B achieves better accuracy. These results demonstrate that when fine-tuning all model parameters with an accuracy-driven SFT, the scale advantage of a large dataset can be better exploited.

D. Disscusion on Efficiency

Although our method utilizes rich features from large datasets, we select moderate-sized pre-train models. In addition, the proposed CDTM is relatively lightweight. Therefore, the proposed methods are comparable to classical models in terms of efficiency and are suitable for deployment in medical applications. Taking an input image size of 224×224, the efficiency analysis experiment results on BreakHis dataset are listed in Table VII.

Specifically, the proposed ConvNeXt-B-CDTM-SFT and ViT-B-CDTM-SFT models exhibit superior performance compared to optimal classical networks, with a 12.05% and 10.16% improvement in F1 score, respectively, and a lower parameter count of 39.04 M and 35.32 M, respectively,

Table VI. Comparison of cross-domain fine-tuning results on the BreakHis dataset under different scales of pre-training data.

Model	Pre-train	Pre (%)	Rec (%)	Acc (%)	F1 (%)
Add CDTM and fine-tune with LP					
ConvNeXt-B	ImageNet-21K	82.98	**79.06**	**85.61**	**80.97**
ConvNeXt-B	LAION-400M	**84.76**	76.87	**85.61**	80.62
ConvNeXt-B	LAION-2B	80.78	75.09	84.17	77.83
ViT-B	ImageNet-21K	80.02	**80.73**	84.89	80.37
ViT-B	LAION-400M	**85.86**	79.33	87.05	**82.47**
ViT-B	LAION-2B	84.31	78.33	**87.41**	81.21
Add CDTM and fine-tune with SFT					
ConvNeXt-B	ImageNet-21K	88.76	87.60	91.01	88.18
ConvNeXt-B	LAION-400M	88.18	**88.63**	91.37	88.41
ConvNeXt-B	LAION-2B	**93.09**	87.14	**92.45**	**90.02**
ViT-B	ImageNet-21K	88.99	88.01	91.37	88.50
ViT-B	LAION-400M	**93.15**	90.89	93.88	**92.01**
ViT-B	LAION-2B	92.74	**91.08**	**94.24**	91.91

Table VII. Comparison of Method efficiency.

Method	Params(M)	Flops(G)	F1(%)	Train
ResNet-50	25.56	4.12	75.54	100%
VGG-16	138.36	15.50	79.86	100%
VGG-19	143.67	19.67	77.14	100%
ConvNeXt-B	88.85	15.42	33.26	100%
ConvNeXt-B-CDTM-LP	99.32	15.88	77.83	10.5%
ConvNeXt-B-CDTM-SFT	99.32	15.88	90.02	100%
ViT-B	85.77	16.86	61.87	100%
ViT-B-CDTM-LP	102.84	17.90	81.21	16.6%
ViT-B-CDTM-SFT	102.84	17.90	91.91	100%

with no significant difference in FLops. Notably, under the efficiency-driven LP strategy, ConvNeXt-B-CDTM-LP and ViT-B-CDTM-LP achieve F1 scores that surpass classical networks, using only 10.5% and 16.6% of their training parameters, respectively. These results highlight the significant advantages of our approach in terms of both precision and efficiency.

In addition, the proposed CDTM Fine-tuning with the SFT strategy yields a remarkable 170.75% improvement in the F1 score for ConvNeXt-B, despite a slight increase of 11.78% in parameter count and 2.98% in Flops. Similarly, ViT-B-CDTM demonstrates a 48.55% improvement in F1 score, despite having a higher parameter count and Flops by 19.90% and 1.04%, respectively, compared to ViT-B. These experimental results demonstrate the significant performance enhancement of the proposed CDTM on medical images through cross-domain transfer, while also considering efficiency.

E. Ablation Study on the Proposed Method

In the *Method* section, we propose and analyze the advantages of the Cross-Domain Transfer Module (CDTM) and Staged Fine-tuning (SFT). To verify the importance of each proposed component, this section shows detailed comparison and ablation studies.

Table VIII lists the results of ablation studies from cross-domain fine-tuning on the LAION-2B dataset to the BreakHis dataset, with training from scratch as the baseline (i.e., without pre-training on any dataset).

Table VIII. Ablation study of proposed components on LAION-2B cross-domain fine-tuning to BreakHis.

Model	Fine-tune	CDTM	Acc (%)	F1 (%)
ConvNeXt-B	From Scratch		65.47	33.26
ConvNeXt-B	LP		72.66	55.97
ConvNeXt-B	LP	✓	**84.17**	**77.83**
ConvNeXt-B	FFT		89.93	86.01
ConvNeXt-B	**SFT**	✓	**92.45**	**90.02**
ViT-B	From Scratch		74.10	61.87
ViT-B	LP		64.75	31.60
ViT-B	LP	✓	**87.41**	**81.21**
ViT-B	FFT		89.93	86.39
ViT-B	**SFT**	✓	**94.24**	**91.91**

The results show that under efficiency-driven LP fine-tuning, the ConvNeXt model outperforms training from scratch on medical datasets. However, the results of ViT are inferior to training from scratch, indicating that ViT models cannot complete cross-domain transfer learning when the backbone is frozen under LP. The proposed CDTM provides a solution to this dilemma. Specifically, With the proposed CDTM module, the cross-domain fine-tuning performance of ConvNeXt improves further, with an F1 score increase of 21.86%. ViT achieves an even more significant improvement, with an F1 score increase of 49.61% compared to training from scratch and 19.34% compared to training without CDTM. These results demonstrate the proposed CDTM provides powerful cross-domain transfer capability, which is highly beneficial for models to transfer rich features from large-scale natural visual datasets to medical images.

The proposed accuracy-driven SFT strategy, combined with CDTM, maximizes the potential of large visual datasets and pre-training models. Table VIII shows that using this SFT, ConvNeXt achieves a 4.01% higher F1 score on the medical dataset than FFT, and ViT achieves a 5.52% higher F1 score. When compared to the efficiency-driven PT, the accuracy-driven SFT results in an average 11.45% higher F1 score. These experimental results demonstrate the effectiveness of the SFT strategy, which further facilitates cross-domain transfer, and collaborates with CDTM to achieve state-of-the-art results for the models on the medical dataset.

V. Conclusions

In this paper, we propose a novel method to cross-domain enhance medical image analysis using pre-trained models on large-scale natural datasets. Our method is capable of transferring rich features from natural vision datasets to medical images using the Cross-Domain Transfer Module (CDTM) and the Stage Fine-Tuning (SFT) strategy. The CDTM enables cross-domain transfer by introducing a cluster center attention mechanism. The SFT combines the characteristics of CDTM and designs the fine-tuning strategy, which further improves the performance of the model on downstream medical datasets. Validation experiments and ablation analysis demonstrate the effectiveness and advantages of our method over state-of-the-art algorithms.

REFERENCES

[1] H. Fu, G. Wang, W. Lei, W. Xu, Q. Zhao, S. Zhang, K. Li, and S. Zhang, "HMRNet: High and multi-resolution network with bidirectional feature calibration for brain structure segmentation in radiotherapy," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 9, pp. 4519–4529, 2022.

[2] J. Chai, H. Zeng, A. Li, and E. W. Ngai, "Deep learning in computer vision: A critical review of emerging techniques and application scenarios," *Machine Learning with Applications*, vol. 6, p. 100134, 2021.

[3] A. Papadopoulos and A. Delopoulos, "Leveraging unlabelled data in multiple-instance learning problems for improved detection of parkinsonian tremor in free-living conditions," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[4] J.-W. Zhang, W. Chen, K. I. Ly, X. Zhang, F. Yan, J. Jordan, G. Harris, S. Plotkin, P. Hao, and W. Cai, "DINs: deep interactive networks for neurofibroma segmentation in neurofibromatosis type 1 on whole-body mri," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 2, pp. 786–797, 2021.

[5] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5B: An open large-scale dataset for training next generation image-text models," *ArXiv Preprint arXiv:2210.08402*, 2022.

[6] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, pp. 211–252, 2015.

[7] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "Laion-400M: Open dataset of clip-filtered 400 million image-text pairs," *ArXiv Preprint arXiv:2111.02114*, 2021.

[8] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4," *ArXiv Preprint arXiv:2303.12712*, 2023.

[9] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE Access*, vol. 8, pp. 58 443–58 469, 2020.

[10] M. A. Al-Garadi, A. Mohamed, A. K. Al-Ali, X. Du, I. Ali, and M. Guizani, "A survey of machine and deep learning methods for internet of things (IoT) security," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1646–1685, 2020.

[11] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[12] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 590–597.

[13] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Jama*, vol. 316, no. 22, pp. 2402–2410, 2016.

[14] L. Alzubaidi, M. Al-Amidie, A. Al-Asadi, A. J. Humaidi, O. Al-Shamma, M. A. Fadhel, J. Zhang, J. Santamaría, and Y. Duan, "Novel transfer learning approach for medical imaging with limited labeled data," *Cancers*, vol. 13, no. 7, p. 1590, 2021.

[15] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler *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 *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 168–172.

[16] 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, no. 1, pp. 1–9, 2018.

[17] M. Combalia, N. C. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig *et al.*, "Bcn20000: Dermoscopic lesions in the wild," *ArXiv Preprint arXiv:1908.02288*, 2019.

[18] X. He, X. Yang, S. Zhang, J. Zhao, Y. Zhang, E. Xing, and P. Xie, "Sample-efficient deep learning for COVID-19 diagnosis based on CT scans," *Medrxiv*, pp. 2020–04, 2020.

[19] Y. Liu, A. Jain, C. Eng, D. H. Way, K. Lee, P. Bui, K. Kanada, G. de Oliveira Marinho, J. Gallegos, S. Gabriele *et al.*, "A deep learning system for differential diagnosis of skin diseases," *Nature Medicine*, vol. 26, no. 6, pp. 900–908, 2020.

[20] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. S. Corrado, A. Darzi *et al.*, "International evaluation of an AI system for breast cancer screening," *Nature*, vol. 577, no. 7788, pp. 89–94, 2020.

[21] B. Neyshabur, H. Sedghi, and C. Zhang, "What is being transferred in transfer learning?" *Advances in Neural Information Processing Systems*, vol. 33, pp. 512–523, 2020.

[22] M. Heker and H. Greenspan, "Joint liver lesion segmentation and classification via transfer learning," *ArXiv Preprint arXiv:2004.12352*, 2020.

[23] F. Spanhol, L. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological database (BreakHis)," 2021.

[24] C. Sun, C. Li, J. Zhang, M. M. Rahaman, S. Ai, H. Chen, F. Kulwa, Y. Li, X. Li, and T. Jiang, "Gastric histopathology image segmentation using a hierarchical conditional random field," *Biocybernetics and Biomedical Engineering*, vol. 40, no. 4, pp. 1535–1555, 2020.

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[26] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[27] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," *ArXiv Preprint arXiv:2106.08254*, 2021.

[28] X. Huang and Q. Li, "Runner-up solution to google universal image embedding competition 2022," *ArXiv Preprint arXiv:2210.08735*, 2022.

[29] L. Zhang, T. Xiang, and S. Gong, "Learning a deep embedding model for zero-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2021–2030.

[30] B. Hanin, "Which neural net architectures give rise to exploding and vanishing gradients?" *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[31] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *ArXiv Preprint arXiv:1607.06450*, 2016.

[32] B. Gao and L. Pavel, "On the properties of the softmax function with application in game theory and reinforcement learning," *ArXiv Preprint arXiv:1704.00805*, 2017.

[33] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," *ArXiv Preprint arXiv:1606.08415*, 2016.

[34] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2009, pp. 248–255.

[35] H. Chen, C. Li, G. Wang, X. Li, M. M. Rahaman, H. Sun, W. Hu, Y. Li, W. Liu, C. Sun *et al.*, "GasHis-Transformer: A multi-scale visual transformer approach for gastric histopathological image detection," *Pattern Recognition*, vol. 130, p. 108827, 2022.

[36] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *ArXiv Preprint arXiv:2010.11929*, 2020.

[37] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.

[38] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[40] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ArXiv Preprint arXiv:1409.1556*, 2014.

[42] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 519–16 529.

[43] Y. Dai, Y. Gao, and F. Liu, "TransMed: Transformers advance multi-modal medical image classification," *Diagnostics*, vol. 11, no. 8, p. 1384, 2021.
[44] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, and M. Douze, "LeVit: a vision transformer in convnet's clothing for faster inference," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 259–12 269.
[45] Y. Li, X. Wu, C. Li, X. Li, H. Chen, C. Sun, M. M. Rahaman, Y. Yao, Y. Zhang, and T. Jiang, "A hierarchical conditional random field-based attention mechanism approach for gastric histopathology image classification," *Applied Intelligence*, pp. 1–22, 2022.
[46] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.
[47] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.
[48] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *ArXiv Preprint arXiv:1810.04805*, 2018.