MDViT: Multi-domain Vision Transformer for Small Medical Image Segmentation Datasets

Siyi Du[1][0000−0002−9961−4533], Nourhan Bayasi[1][0000−0003−4653−6081], Ghassan Hamarneh[2][0000−0001−5040−7448], and Rafeef Garbi[1][0000−0001−6224−0876]

[1] University of British Columbia, Vancouver, British Columbia, CA
{siyi,nourhanb,rafeef}@ece.ubc.ca
[2] Simon Fraser University, Burnaby, British Columbia, CA
hamarneh@sfu.ca

Abstract. Despite its clinical utility, medical image segmentation (MIS) remains a daunting task due to images' inherent complexity and variability. Vision transformers (ViTs) have recently emerged as a promising solution to improve MIS; however, they require larger training datasets than convolutional neural networks. To overcome this obstacle, data-efficient ViTs were proposed, but they are typically trained using a single source of data, which overlooks the valuable knowledge that could be leveraged from other available datasets. Naïvly combining datasets from different domains can result in negative knowledge transfer (NKT), i.e., a decrease in model performance on some domains with non-negligible inter-domain heterogeneity. In this paper, we propose MDViT, the first multi-domain ViT that includes domain adapters to mitigate data-hunger and combat NKT by adaptively exploiting knowledge in multiple small data resources (domains). Further, to enhance representation learning across domains, we integrate a mutual knowledge distillation paradigm that transfers knowledge between a universal network (spanning all the domains) and auxiliary domain-specific network branches. Experiments on 4 skin lesion segmentation datasets show that MDViT outperforms state-of-the-art algorithms, with superior segmentation performance and a fixed model size, at inference time, even as more domains are added. Our code is available at https://github.com/siyi-wind/MDViT.

Keywords: Vision Transformer · Data-efficiency · Multi-domain Learning · Medical Image Segmentation · Dermatology.

1 Introduction

Medical image segmentation (MIS) is a crucial component in medical image analysis, which aims to partition an image into distinct regions (or segments) that are semantically related and/or visually similar. This process is essential for clinicians to, among others, perform qualitative and quantitative assessments of various anatomical structures or pathological conditions and perform image-guided treatments or treatment planning [2]. Vision transformers (ViTs), with their inherent ability to model long-range dependencies, have recently been considered a

promising technique to tackle MIS. They process images as sequences of patches, with each patch having a global view of the entire image. This enables a ViT to achieve improved segmentation performance compared to traditional convolutional neural networks (CNNs) on plenty of segmentation tasks [16]. However, due to the lack of inductive biases, such as weight sharing and locality, ViTs are more data-hungry than CNNs, i.e., require more data to train [31]. Meanwhile, it is common to have access to multiple, diverse, yet small-sized datasets (100s to 1000s of images per dataset) for the same MIS task, e.g., PH2 [25] and ISIC 2018 [11] in dermatology, LiTS [6] and CHAOS [18] in liver CT, or OASIS [24] and ADNI [17] in brain MRI. As each dataset alone is too small to properly train a ViT, the challenge becomes how to effectively leverage the different datasets.

Table 1: Related works on mitigating ViTs' data-hunger or multi-domain adaptive learning. **U** (universal) implies a model spans multiple domains. **F** means the model's size at inference time remains fixed even when more domains are added.

Method	**ViT**	**Mitigate ViTs' data-hunger**	**U**	**F**
[7,39,22]	√	√ by adding inductive bias	×	-
[31,37]	√	√ by knowledge sharing	×	-
[34]	√	√ by increasing dataset size	×	-
[8]	√	√ by unsupervised pretraining	×	-
[28,35]	×	×	√	√
[21,26]	×	×	√	×
[32]	√	×	√	×
MDViT	√	√ by multi-domain learning	√	√

Various strategies have been proposed to address ViTs' data-hunger (Table 1), mainly: *Adding inductive bias* by constructing a hybrid network that fuses a CNN with a ViT [39], imitating CNNs' shifted filters and convolutional operations [7], or enhancing spatial information learning [22]; *sharing knowledge* by transferring knowledge from a CNN [31] or pertaining ViTs on multiple related tasks and then fine-tuning on a down-stream task [37]; *increasing data* via augmentation [34]; and *non-supervised pre-training* [8]. Nevertheless, one notable limitation in these approaches is that they are not universal, i.e., they rely on *separate training* for each dataset rather than incorporate valuable knowledge from related domains. As a result, they can incur additional training, inference, and memory costs, which is especially challenging when dealing with multiple small datasets in the context of MIS tasks. Multi-domain learning, which trains a single universal model to tackle all the datasets simultaneously, has been found promising for reducing computational demands while still leveraging information from multiple domains [1,21]. To the best of our knowledge, multi-domain universal models have not yet been investigated for alleviating ViTs' data-hunger.

Given the inter-domain heterogeneity resulting from variations in imaging protocols, scanner manufacturers, etc. [4,21], directly mixing all the datasets for training, i.e., *joint training*, may improve a model's performance on one dataset while degrading performance on other datasets with non-negligible unrelated domain-specific information, a phenomenon referred to as *negative knowledge transfer* (NKT) [1,38]. A common strategy to mitigate NKT in computer vision is to introduce adapters aiding the model to adapt to different domains, i.e., *multi-domain adaptive training* (MAT), such as domain-specific mechanisms [21,26,32],

and squeeze-excitation layers [35,28] (Table 1). However, those MAT techniques are built based on CNN rather than ViT or are scalable, i.e., the models' size at the inference time increases linearly with the number of domains.

To address ViTs' data-hunger, in this work, we propose MDViT, a novel fixed-size multi-domain ViT trained to adaptively aggregate valuable knowledge from multiple datasets (domains) for improved segmentation. In particular, we introduce a domain adapter that adapts the model to different domains to mitigate negative knowledge transfer caused by inter-domain heterogeneity. Besides, for better representation learning across domains, we propose a novel mutual knowledge distillation approach that transfers knowledge between a universal network (spanning all the domains) and additional domain-specific network branches.

We summarize our contributions as follows: (1) To the best of our knowledge, we are the first to introduce multi-domain learning to alleviate ViTs' data-hunger when facing limited samples per dataset. (2) We propose a multi-domain ViT, MDViT, for medical image segmentation with a novel domain adapter to counteract negative knowledge transfer and with mutual knowledge distillation to enhance representation learning. (3) The experiments on 4 skin lesion segmentation datasets show that our multi-domain adaptive training outperforms separate and joint training (ST and JT), especially a 10.16% improvement in IOU on the skin cancer detection dataset compared to ST and that MDViT outperforms state-of-the-art data-efficient ViTs and multi-domain learning strategies.

2 Methodology

Let $\boldsymbol{X} \in \mathbb{R}^{H\times W\times 3}$ be an input RGB image and $\boldsymbol{Y} \in \{0,1\}^{H\times W}$ be its ground-truth segmentation mask. Training samples $\{(\boldsymbol{X},\boldsymbol{Y})\}$ come from M datasets, each representing a domain. We aim to build and train a single ViT that performs well on all domain data and addresses the insufficiency of samples in any of the datasets. We first introduce our baseline (BASE), a ViT with hierarchical transformer blocks (Fig. 1-a). Our proposed MDViT extends BASE with 1) a domain adapter (DA) module inside the factorized multi-head self-attention (MHSA) to adapt the model to different domains (Fig. 1-b,c), and 2) a mutual knowledge distillation (MKD) strategy to extract more robust representations across domains (Fig. 1-d). We present the details of MDViT in Section 2.1.

BASE is a U-shaped ViT based on the architecture of U-Net [27] and pyramid ViTs [19,7]. It contains encoding (the first four) and decoding (the last four) transformer blocks, a two-layer CNN bridge, and skip connections. As described in [19], the ith transformer block involves a convolutional patch embedding layer with a patch size of 3×3 and L_i transformer layers with factorized MHSA in linear complexity, the former of which converts a feature map X_{i-1} into a sequence of patch embeddings $\boldsymbol{z}_i \in \mathbb{R}^{N_i\times C_i}$, where $N_i = \frac{H}{2^{i+1}}\frac{W}{2^{i+1}}, 1 \le i \le 4$ is the number of patches and C_i is the channel dimension. We use the same position embedding as [19] and skip connections as [27]. To reduce computational complexity, following [19], we add two and one CNN layer before and after transformer blocks, respectively, enabling the 1st transformer block to process features starting from

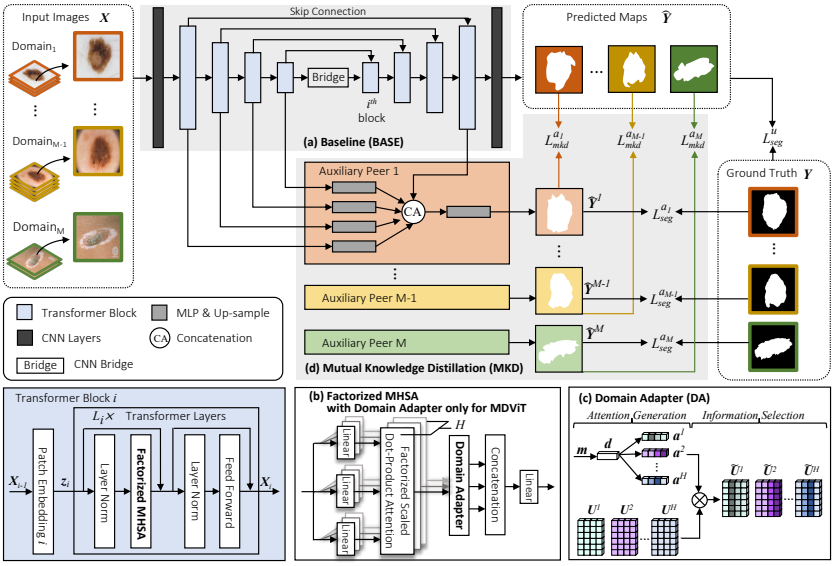

Fig. 1: Overall architecture of MDViT, which is trained on multi-domain data by optimizing two types of losses: L_{seg} and L_{mkd}. MDViT extends BASE (a) with DA inside factorized MHSA (b), which is detailed in (c), and MKD (d).

a lower resolution: $\frac{H}{4} \times \frac{W}{4}$. We do not employ integrated and hierarchical CNN backbones, e.g., ResNet, in BASE as data-efficient hybrid ViTs [33,39], to clearly evaluate the efficacy of multi-domain learning in mitigating ViTs' data-hunger.

2.1 MDViT

MDViT consists of a universal network (spanning M domains) and M auxiliary network branches, i.e., peers, each associated with one of the M domains. The universal network is the same as BASE, except we insert a domain adapter (DA) in each factorized MHSA to tackle negative knowledge transfer. Further, we employ a mutual knowledge distillation (MKD) strategy to transfer domain-specific and shared knowledge between peers and the universal network to enhance representation learning. Next, we will introduce DA and MKD in detail.

Domain Adapter (DA): In multi-domain adaptive training, some methods build domain-specific layers in parallel with the main network [26,21,32]. Without adding domain-specific layers, we utilize the existing parallel structure in ViTs, i.e., MHSA, for domain adaptation. The H parallel heads of MHSA mimic how humans examine the same object from different perspectives [10]. Similarly, our intuition of inserting the DA into MHSA is to enable the different heads to have varied perspectives across domains. Rather than manually designate each head to one of the domains, guided by a domain label, MDViT learns to focus on the corresponding features from different heads when encountering a domain. DA contains two steps: *Attention Generation* and *Information Selection* (Fig. 1-c).

Attention Generation generates attention for each head. We first pass a domain label vector $\boldsymbol{m}$ (we adopt one-hot encoding $\boldsymbol{m} \in \mathbb{R}^M$ but other encodings are possible) through one linear layer with a ReLU activation function to acquire a domain-aware vector $\boldsymbol{d} \in \mathbb{R}^{\frac{K}{r}}$. K is the channel dimension of features from the heads. We set the reduction ratio r to 2. After that, similar to [20], we calculate attention for each head: $\boldsymbol{a}^h = \psi(\boldsymbol{W}^h \boldsymbol{d}) \in \mathbb{R}^K, h = 1, 2, ...H$, where ψ is a softmax operation across heads and $\boldsymbol{W}^h \in \mathbb{R}^{K \times \frac{K}{r}}$.

Information Selection adaptively selects information from different heads. After getting the feature $\boldsymbol{U}^h = [\boldsymbol{u}_1^h, \boldsymbol{u}_2^h, ..., \boldsymbol{u}_K^h] \in \mathbb{R}^{N \times K}$ from the hth head, we utilize $\boldsymbol{a}^h$ to calibrate the information along the channel dimension: $\tilde{\boldsymbol{u}}_k^h = a_k^h \cdot \boldsymbol{u}_k^h$.
Mutual Knowledge Distillation (MKD): Distilling knowledge from domain-specific networks has been found beneficial for universal networks to learn more robust representations [21,40]. Moreover, mutual learning that transfers knowledge between teachers and students enables both to be optimized simultaneously [15]. To realize these benefits, we propose MKD that mutually transfers knowledge between auxiliary peers and the universal network. In Fig. 1-d, the mth auxiliary peer is only trained on the mth domain, producing output $\hat{\boldsymbol{Y}}^m$, whereas the universal network's output is $\hat{\boldsymbol{Y}}$. Similar to [21], we utilize a symmetric Dice loss $L_{mkd}^{a_m} = Dice(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{Y}}^m)$ as the knowledge distillation loss. Each peer is an expert in a certain domain, guiding the universal network to learn domain-specific information. The universal network experiences all the domains and grasps the domain-shared knowledge, which is beneficial for peer learning.

Each *Auxiliary Peer* is trained on a small, individual dataset specific to that peer (Fig. 1-d). To achieve a rapid training process and prevent overfitting, particularly when working with numerous training datasets, we adapt a lightweight multilayer perception (MLP) decoder designed for ViT encoders [36] to our peers' architecture. Specifically, multi-level features from the encoding transformer blocks (Fig. 1-a) go through an MLP layer and an up-sample operation to unify the channel dimension and resolution to $\frac{H}{4} \times \frac{W}{4}$, which are then concatenated with the feature involving domain-shared information from the universal network's last transformer block. Finally, we pass the fused feature to an MLP layer and do an up-sample to obtain a segmentation map.

2.2 Objective Function

Similar to Combo loss [29], BASE's segmentation loss combines Dice and binary cross entropy loss: $L_{seg} = L_{Dice} + L_{bce}$. In MDViT, we use the same segmentation loss for the universal network and auxiliary peers, denoted as L_{seg}^u and L_{seg}^a, respectively. The overall loss is calculated as follows.

$$L_{total} = L_{seg}^u(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) + \alpha \sum_{m=1}^{M} L_{seg}^{a_m}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}^m) + \beta \sum_{m=1}^{M} L_{mkd}^{a_m}(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{Y}}^m). \quad (1)$$

We set both α and β to 0.5. $L_{seg}^{a_m}$ does not optimize DA to avoid interfering with the domain adaptation learning. After training, we discard the auxiliary peers and only utilize the universal network for inference.

Table 2: Segmentation results comparing BASE, MDViT, and SOTA methods. We report the models' parameter count at inference time in millions (M). **T** means training paradigms. † represents using domain-specific normalization.

Model	#Param. (millions) (M)	T	Segmentation Results in Test Sets (%)									
			Dice ↑					IOU ↑				
			ISIC	DMF	SCD	PH2	avg ± std	ISIC	DMF	SCD	PH2	avg ± std
(a) BASE												
BASE	27.8×	ST	90.18	90.68	86.82	93.41	90.27 ± 1.16	82.82	83.22	77.64	87.84	82.88 ± 1.67
BASE	27.8	JT	89.42	89.89	92.96	94.24	91.63 ± 0.42	81.68	82.07	87.03	89.36	85.04 ± 0.64
(b) Our Method												
MDViT	28.5	MAT	90.29	90.78	**93.22**	**95.53**	**92.45 ± 0.65**	82.99	83.41	**87.80**	**91.57**	**86.44 ± 0.94**
(c) Other Data-efficient MIS ViTs												
SwinUnet	41.4×	ST	89.25	90.69	88.58	94.13	90.66 ± 0.87	81.51	83.25	80.40	89.00	83.54 ± 1.27
SwinUnet	41.4	JT	89.64	90.40	92.98	94.86	91.97 ± 0.30	81.98	82.80	87.08	90.33	85.55 ± 0.50
UTNet	10.0×	ST	89.74	90.01	88.13	93.23	90.28 ± 0.62	82.16	82.13	79.87	87.60	82.94 ± 0.82
UTNet	10.0	JT	90.24	89.85	92.06	94.75	91.72 ± 0.63	82.92	82.00	85.66	90.17	85.19 ± 0.96
BAT	32.2×	ST	**90.45**	90.56	90.78	94.72	91.63 ± 0.68	83.04	82.97	83.66	90.03	84.92 ± 1.01
BAT	32.2	JT	90.06	90.06	92.66	93.53	91.58 ± 0.33	82.44	82.18	86.48	88.11	84.80 ± 0.53
TransFuse	26.3×	ST	90.43	**91.04**	91.37	94.93	91.94 ± 0.67	**83.18**	**83.86**	84.91	90.44	85.60 ± 0.95
TransFuse	26.3	JT	90.03	90.48	92.54	95.14	92.05 ± 0.36	82.56	82.97	86.50	90.85	85.72 ± 0.56
Swin UNETR	25.1×	ST	90.29	90.95	91.10	94.45	91.70 ± 0.51	82.93	83.69	84.16	89.59	85.09 ± 0.79
Swin UNETR	25.1	JT	89.81	90.87	92.29	94.73	91.93 ± 0.29	82.21	83.58	86.10	90.11	85.50 ± 0.44
(d) Other Multi-domain Learning Methods												
Rundo et al.	28.2	MAT	89.43	89.46	92.62	94.68	91.55 ± 0.64	81.73	81.40	86.71	90.12	84.99 ± 0.90
Wang et al.	28.1	MAT	89.46	89.62	92.62	94.47	91.55 ± 0.54	81.79	81.59	86.71	89.76	84.96 ± 0.74
BASE†	27.8(.02×)	MAT	90.22	90.61	**93.69**	95.55	92.52 ± 0.45	82.91	83.14	**88.28**	91.58	86.48 ± 0.74
MDViT†	28.6(.02×)	MAT	**90.24**	**90.71**	93.38	**95.90**	**92.56 ± 0.52**	**82.97**	**83.31**	88.06	**92.19**	**86.64 ± 0.76**

3 Experiments

Datasets and Evaluation Metrics: We study 4 skin lesion segmentation datasets collected from varied sources: ISIC 2018 (ISIC) [11], Dermofit Image Library (DMF) [3], Skin Cancer Detection (SCD) [14], and PH2 [25], which contain 2594, 1300, 206, and 200 samples, respectively. To facilitate a fairer performance comparison across datasets, as in [4], we only use the 1212 images from DMF that exhibited similar lesion conditions as those in other datasets. We perform 5-fold cross-validation and utilize Dice and IOU metrics for evaluation as [33].
Implementation Details: We conduct 3 training paradigms: separate (ST), joint (JT), and multi-domain adaptive training (MAT), described in Section 1, to train all the models from scratch on the skin datasets. Images are resized to 256×256 and then augmented through random scaling, shifting, rotation, flipping, Gaussian noise, and brightness and contrast changes. The encoding transformer blocks' channel dimensions are [64, 128, 320, 512] (Fig. 1-a). We use two transformer layers in each transformer block and set the number of heads in MHSA to 8. The hidden dimensions of the CNN bridge and auxiliary peers are 1024 and 512. We deploy models on a single TITAN V GPU and train them for 200 epochs with the AdamW [23] optimizer, a batch size of 16, ensuring 4 samples from each dataset, and an initial learning rate of 1×10^{-4}, which changes through a linear decay scheduler whose step size is 50 and decay factor $\gamma = 0.5$.
Comparing Against BASE: In Table 2-a,b, compared with BASE in ST, BASE in JT improves the segmentation performance on small datasets (PH2 and SCD) but at the expense of diminished performance on larger datasets (ISIC and DMF). This is expected given the non-negligible inter-domain heterogeneity between skin lesion datasets, as found by Bayasi et al. [5]. The above results

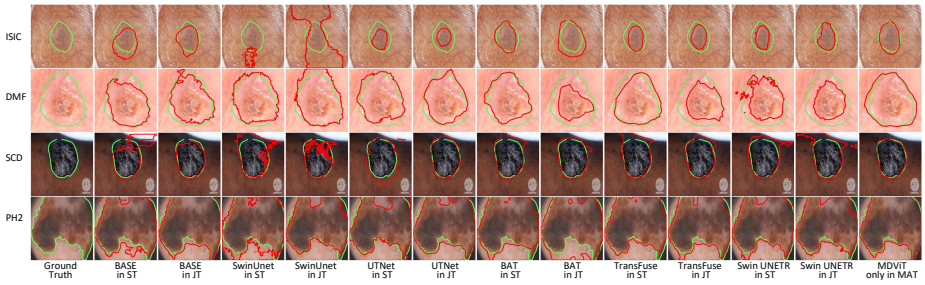

Fig. 2: Visual result comparison of MDViT, BASE and SOTA data-efficient MIS ViTs in ST and JT training paradigms on four datasets. The green and red contours present the ground truth and segmentation results, respectively.

demonstrate that shared knowledge in related domains facilitates training a ViT on small datasets while, without a well-designed multi-domain algorithm, causing negative knowledge transfer (NKT) due to inter-domain heterogeneity, i.e., the model's performance decreases on other datasets. Meanwhile, MDViT fits all the domains without NKT and outperforms BASE in ST by a large margin; significantly increasing Dice and IOU on SCD by 6.4% and 10.16%, showing that MDViT smartly selects valuable knowledge when given data from a certain domain. Additionally, MDViT outperforms BASE in JT across all the domains, with average improvements of 0.82% on Dice and 1.4% on IOU.

Comparing Against State-of-the-Art (SOTA) Methods: We conduct experiments on SOTA data-efficient MIS ViTs and multi-domain learning methods. Previous MIS ViTs mitigated the data-hunger in one dataset by adding inductive bias, e.g., SwinUnet [7], UTNet [13], BAT [33], TransFuse [39], and Swin UNETR [30]. We implement ResNet-34 as the backbone of BAT for fair comparison (similar model size). As illustrated in Table 2-a,b,c, these SOTA models are superior to BASE in SJ. This is expected since they are designed to reduce data requirements. Nevertheless, in JT, these models also suffer from NKT: They perform better than models in ST on some datasets, like SCD, and worse on others, like ISIC. Finally, MDViT achieves the best segmentation performance in average Dice and IOU without NKT and has the best results on SCD and PH2. Fig. 2 shows MDViT's excellent performance on ISIC and DMF and that it achieves the closest results to ground truth on SCD and PH2. More segmentation results are presented in the supplementary material. Though BAT and TransFuse in ST have better results on some datasets like ISIC, they require extra compute resources to train M models as well as an M-fold increase in memory requirements. The above results indicate that domain-shared knowledge is especially beneficial for training relatively small datasets such as SCD.

We employ the two fixed-size (i.e., independent of M) multi-domain algorithms proposed by Rundo et al. [28] and Wang et al. [35] on BASE. We set the number of parallel SE adapters in [35] to 4. In Table 2-b,d, MDViT outperforms both of them on all the domains, showing the efficacy of MDViT and that

Table 3: Ablation studies of MDViT and experiments of DA's plug-in capability. KD means general knowledge distillation, i.e., we only transfer knowledge from auxiliary peers to the universal network. ^{D} or ^{B} refers to using DeepLabv3's decoder or BASE's decoding layers as auxiliary peers.

Model	**#Param. (M)**	**T**	Dice ↑					IOU ↑				
			ISIC	DMF	SCD	PH2	avg ± std	ISIC	DMF	SCD	PH2	avg ± std
(a) Plug-in Capability of DA												
DosViT	14.6	JT	88.66	89.72	90.65	94.26	90.82 ± 0.43	80.45	81.73	83.29	89.26	83.68 ± 0.68
DosViT+DA	14.9	MAT	89.22	89.91	90.73	94.42	91.07 ± 0.32	81.28	82.00	83.44	89.57	84.07 ± 0.50
TransFuse	26.3	JT	90.03	90.48	92.54	95.14	92.05 ± 0.36	82.56	82.97	86.50	90.85	85.72 ± 0.56
TransFuse+DA	26.9	MAT	90.13	90.47	**93.62**	95.21	92.36 ± 0.38	82.80	82.94	**88.16**	90.97	86.22 ± 0.64
(b) Ablation Study for DA and MKD												
BASE	27.8	JT	89.42	89.89	92.96	94.24	91.63 ± 0.42	81.68	82.07	87.03	89.36	85.04 ± 0.64
BASE+DA	28.5	MAT	89.96	90.66	93.36	95.46	92.36 ± 0.51	82.52	83.24	87.98	91.43	86.29 ± 0.72
BASE+MKD	27.8	JT	89.27	89.53	92.66	94.83	91.57 ± 0.53	81.45	81.49	86.81	90.42	85.04 ± 0.74
BASE+DA+KD	28.5	MAT	90.03	90.59	93.26	**95.63**	92.38 ± 0.39	82.67	83.12	87.85	**91.72**	86.34 ± 0.51
(c) Ablation Study for Auxiliary Peers												
MDViT^{D}	28.5	MAT	89.64	90.25	92.24	95.36	91.87 ± 0.45	82.10	82.55	86.12	91.24	85.50 ± 0.67
MDViT^{B}	28.5	MAT	90.03	90.73	92.72	95.32	92.20 ± 0.50	82.66	83.35	87.01	91.17	86.05 ± 0.70
MDViT	28.5	MAT	**90.29**	**90.78**	93.22	95.53	**92.45 ± 0.65**	**82.99**	**83.41**	87.80	91.57	**86.44 ± 0.94**

multi-domain methods built on ViTs might not perform as well as on CNNs. We also apply the domain-specific normalization [21] to BASE and MDViT to get BASE† and MDViT†, respectively. In Table 2-d, BASE† confronts NKT, which lowers the performance on DMF compared with BASE in ST, whereas MDViT† not only addresses NKT but also outperforms BASE† on average Dice and IOU.

Ablation Studies and Plug-in Capability of DA: We conduct ablation studies to demonstrate the efficacy of DA, MKD, and auxiliary peers. Table 3-b reveals that using one-direction knowledge distillation (KD) or either of the critical components in MDViT, i.e., DA or MKD, but not together, could not achieve the best results. Table 3-c exemplifies that, for building the auxiliary peers, our proposed MLP architecture is more effective and has fewer parameters (1.6M) than DeepLabv3's decoder [9] (4.7M) or BASE's decoding layers (10.8M). Finally, we incorporate DA into two ViTs: TransFuse and DosViT (the latter includes the earliest ViT encoder [12] and a DeepLabv3's decoder). As shown in Table 3-a,b, DA can be used in various ViTs but is more advantageous in MDViT with more transformer blocks in the encoding and decoding process.

4 Conclusion

We propose a new algorithm to alleviate vision transformers (ViTs)' data-hunger in small datasets by aggregating valuable knowledge from multiple related domains. We constructed MDViT, a robust multi-domain ViT leveraging novel domain adapters (DAs) for negative knowledge transfer mitigation and mutual knowledge distillation (MKD) for better representation learning. MDViT is non-scalable, i.e., has a fixed model size at inference time even as more domains are added. The experiments on 4 skin lesion segmentation datasets show that MDViT outperformed SOTA data-efficient medical image segmentation ViTs and multi-domain learning methods. Our ablation studies and application of DA on other ViTs show the effectiveness of DA and MKD and DA's plug-in capability.

References

1. Adadi, A.: A survey on data-efficient algorithms in big data era. Journal of Big Data **8**(1), 24 (2021)
2. Asgari Taghanaki, S., Abhishek, K., Cohen, J.P., Cohen-Adad, J., Hamarneh, G.: Deep semantic segmentation of natural and medical images: a review. Artificial Intelligence Review **54**, 137–178 (2021)
3. Ballerini, L., Fisher, R.B., Aldridge, B., Rees, J.: A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions. In: Color medical image analysis, pp. 63–86. Springer (2013)
4. Bayasi, N., Hamarneh, G., Garbi, R.: Culprit-Prune-Net: Efficient continual sequential multi-domain learning with application to skin lesion classification. In: MICCAI 2021. pp. 165–175. Springer (2021)
5. Bayasi, N., Hamarneh, G., Garbi, R.: BoosterNet: Improving domain generalization of deep neural nets using culpability-ranked features. In: CVPR 2022. pp. 538–548 (2022)
6. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al.: The liver tumor segmentation benchmark (LiTS). Medical Image Analysis **84**, 102680 (2023)
7. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-Unet: Unet-like pure transformer for medical image segmentation. In: ECCV 2022 Workshops. vol. 13803, pp. 205–218. Springer (2023)
8. Cao, Y.H., Yu, H., Wu, J.: Training vision transformers with only 2040 images. arXiv preprint arXiv:2201.10728 (2022)
9. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
10. Clark, K., Khandelwal, U., Levy, O., Manning, C.D.: What does BERT look at? an analysis of BERT's attention. ACL 2019 p. 276 (2019)
11. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC). arXiv preprint arXiv:1902.03368 (2019)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR 2020 (2020)
13. Gao, Y., Zhou, M., Metaxas, D.N.: UTNet: a hybrid transformer architecture for medical image segmentation. In: MICCAI 2021. pp. 61–71. Springer (2021)
14. Glaister, J., Amelard, R., Wong, A., Clausi, D.A.: MSIM: Multistage illumination modeling of dermatological photographs for illumination-corrected skin lesion analysis. IEEE Transactions on Biomedical Engineering **60**(7), 1873–1883 (2013)
15. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision **129**, 1789–1819 (2021)
16. Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., et al.: A survey on vision transformer. IEEE transactions on pattern analysis and machine intelligence (2022)
17. Jack Jr, C.R., et al.: The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. Journal of Magnetic Resonance Imaging **27**(4), 685–691 (2008)
18. Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., et al.: CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation. Medical Image Analysis **69**, 101950 (2021)
19. Lee, Y., Kim, J., Willette, J., Hwang, S.J.: MPViT: Multi-path vision transformer for dense prediction. In: CVPR 2022. pp. 7287–7296 (2022)

20. Li, X., Wang, W., et al.: Selective kernel networks. In: CVPR 2019. pp. 510–519 (2019)
21. Liu, Q., Dou, Q., Yu, L., Heng, P.A.: MS-Net: multi-site network for improving prostate segmentation with heterogeneous MRI data. IEEE transactions on medical imaging **39**(9), 2713–2724 (2020)
22. Liu, Y., Sangineto, E., Bi, W., Sebe, N., Lepri, B., Nadai, M.: Efficient training of visual transformers with small datasets. NeurIPS 2021 **34**, 23818–23830 (2021)
23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
24. Marcus, D.S., Wang, T.H., Parker, J., et al.: Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. Journal of cognitive neuroscience **19**(9), 1498–1507 (2007)
25. Mendonça, T., Ferreira, P.M., Marques, J.S., Marcal, A.R., Rozeira, J.: PH 2-A dermoscopic image database for research and benchmarking. In: EMBC 2013. pp. 5437–5440. IEEE (2013)
26. Rebuffi, S.A., Bilen, H., Vedaldi, A.: Efficient parametrization of multi-domain deep neural networks. In: CVPR 2018. pp. 8119–8127 (2018)
27. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI 2015. pp. 234–241. Springer (2015)
28. Rundo, L., Han, C., Nagano, Y., Zhang, J., Hataya, R., Militello, C., Tangherloni, A., Nobile, M.S., Ferretti, C., Besozzi, D., et al.: USE-Net: Incorporating squeeze-and-excitation blocks into u-net for prostate zonal segmentation of multi-institutional mri datasets. Neurocomputing **365**, 31–43 (2019)
29. Taghanaki, S.A., Zheng, Y., Zhou, S.K., Georgescu, B., Sharma, P., Xu, D., et al.: Combo loss: Handling input and output imbalance in multi-organ segmentation. Computerized Medical Imaging and Graphics **75**, 24–33 (2019)
30. Tang, Y., Yang, D., Li, W., Roth, H.R., Landman, B., Xu, D., Nath, V., Hatamizadeh, A.: Self-supervised pre-training of swin transformers for 3d medical image analysis. In: CVPR 2022. pp. 20730–20740 (2022)
31. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: ICML 2021. pp. 10347–10357. PMLR (2021)
32. Wallingford, M., Li, H., Achille, A., Ravichandran, A., et al.: Task adaptive parameter sharing for multi-task learning. In: CVPR 2022. pp. 7561–7570 (2022)
33. Wang, J., Wei, L., Wang, L., et al.: Boundary-aware transformers for skin lesion segmentation. In: MICCAI 2021. pp. 206–216. Springer (2021)
34. Wang, W., Zhang, J., Cao, Y., Shen, Y., Tao, D.: Towards data-efficient detection transformers. In: ECCV 2022. pp. 88–105. Springer (2022)
35. Wang, X., Cai, Z., Gao, D., Vasconcelos, N.: Towards universal object detection by domain attention. In: CVPR 2019. pp. 7289–7298 (2019)
36. Xie, E., Wang, W., Yu, Z., et al.: SegFormer: Simple and efficient design for semantic segmentation with transformers. NeurIPS 2021 **34**, 12077–12090 (2021)
37. Xie, Y., Zhang, J., et al.: UniMiSS: Universal medical self-supervised learning via breaking dimensionality barrier. In: ECCV 2022. pp. 558–575. Springer (2022)
38. Zhang, W., Deng, L., Zhang, L., Wu, D.: A survey on negative transfer. IEEE/CAA Journal of Automatica Sinica (2022)
39. Zhang, Y., Liu, H., Hu, Q.: TransFuse: Fusing transformers and CNNs for medical image segmentation. In: MICCAI 2021. pp. 14–24. Springer (2021)
40. Zhou, C., Wang, Z., He, S., Zhang, H., Su, J.: A novel multi-domain machine reading comprehension model with domain interference mitigation. Neurocomputing **500**, 791–798 (2022)