AViT: Adapting Vision Transformers for Small Skin Lesion 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. Skin lesion segmentation (SLS) plays an important role in skin lesion analysis. Vision transformers (ViTs) are considered an auspicious solution for SLS, but they require more training data compared to convolutional neural networks (CNNs) due to their inherent parameter-heavy structure and lack of some inductive biases. To alleviate this issue, current approaches fine-tune pre-trained ViT backbones on SLS datasets, aiming to leverage the knowledge learned from a larger set of natural images to lower the amount of skin training data needed. However, fully fine-tuning all parameters of large backbones is computationally expensive and memory intensive. In this paper, we propose AViT, a novel efficient strategy to mitigate ViTs' data-hunger by transferring any pre-trained ViTs to the SLS task. Specifically, we integrate lightweight modules (adapters) within the transformer layers, which modulate the feature representation of a ViT without updating its pre-trained weights. In addition, we employ a shallow CNN as a prompt generator to create a prompt embedding from the input image, which grasps fine-grained information and CNN's inductive biases to guide the segmentation task on small datasets. Our quantitative experiments on 4 skin lesion datasets demonstrate that AViT achieves competitive, and at times superior, performance to SOTA but with significantly fewer trainable parameters. Our code is available at https://github.com/siyi-wind/AViT.

Keywords: Vision Transformer · Data-efficiency · Efficiency · Medical Image Segmentation · Dermatology.

1 Introduction

Melanoma is the most common and dangerous skin malignancy estimated to cause 97,610 new cases and 7,990 deaths in 2023 the United States alone [32], yet early diagnosis and treatment are highly likely to cure it. Automated skin lesion segmentation (SLS), which provides thorough qualitative and quantitative information such as location and border, is a challenging and fundamental operation in computer-aided diagnosis [30]. As a pre-processing step of diagnosis, it boosts the accuracy and robustness of classification by regularizing attention

maps [40], offering the region of interest for wide-field images [4], or removing lesion-adjacent confounding artifacts [27,1]. On the other hand, SLS can serve as a simultaneously optimizing task for classification, enabling the models to obtain improved performance on both two tasks [39]. SLS is also essential for skin color fairness research [12], where the segmented non-lesion area is used to approximate skin tone [22]. Vision transformers (ViTs), with their inherent capability to model global image context through the self-attention mechanism, are a set of promising tools to tackle SLS [17]. Though ViTs have shown improved performance compared to traditional convolutional neural networks (CNNs) [24], they are more data-hungry than CNNs, i.e., need more training data, given the lack of some useful inductive biases like weight sharing and locality [35]. This poses a significant challenge in SLS due to the limited availability of training images, where datasets often contain only a few hundred [29] or thousand [8] samples.

To alleviate ViTs' data-hunger, previous SLS works incorporated some inductive biases through hierarchical architecture [5], local self-attention [34], or convolution layers [14]. Nevertheless, they trained the models from scratch and overlooked the potential benefits of pre-trained models and valuable information from other domains with abundant data. As transfer learning from ImageNet [9] has been demonstrated advantageous for skin lesion tasks [28], an increasingly popular and promising way is to deploy a large pre-trained ViT as the encoder, and then fine-tune the entire model [37,42]. Despite achieving better performance, these techniques that rely on transfer learning have two notable drawbacks. First, a robust ViT typically has plenty of parameters, e.g., ViT-Base (86 million (M)) [10] and Swin-Base (88M) [25], thus making the full fine-tuning strategy quite expensive in terms of computation and memory requirements, especially when dealing with multiple datasets, i.e., we need to store an entire model for each dataset. Second, updating all parameters of a large-scale pre-trained model (full fine-tuning) on smaller datasets is found to be unstable [31] and may instead undermine the model's generalizable representations [41].

The newer parameter-efficient fine-tuning (PEFT) has been proposed as an effective and efficient solution, which only tunes a small subset of the model's parameters. PEFT in computer vision can be divided into two main directions: 1) prompt tuning [21,2] and 2) adapter tuning [41,38,7]. The first direction uses soft (i.e., tunable) prompts: task-specific parameters introduced into the frozen pre-trained ViT backbone's input space and tuned throughout the task-learning process. For example, Jia et al. [21] utilized randomly initialized trainable parameters as soft prompts and prepended them to pre-trained ViT's input for downstream recognition tasks. The second direction uses adapters: trainable lightweight modules inserted into the transformer layers, to modify the hidden representation of the frozen ViT rendering it suitable for a specific task. These PEFT approaches have shown substantially increased efficiency with comparable, or even improved, performance compared to those of full fine-tuning on low-data regimes. Nonetheless, very few works have adapted PEFT to medical imaging. Wu et al. [38] employed adapters to steer the Segment Anything Model (SAM) [23], a promptable ViT-based foundation model trained using 1 billion

masks, to medical image segmentation tasks without updating SAM's parameters. However, they require additional pre-training on medical imaging data prior to adaptation as well as hard prompts in the form of un-tunable information input, such as free-form text or a set of foreground/background points, which increases the computational cost and necessitates prior information collection.

To address ViTs' data-hunger while maintaining the model's efficiency, in this work, we propose AViT, a novel transfer learning strategy that adapts a pre-trained ViT backbone to small SLS datasets by using PEFT. We incorporate lightweight adapter modules into the transformer layers to modify the image representation and keep the pre-trained weights untouched. Furthermore, to enhance the information extraction, we introduce a shallow CNN network in parallel with ViT as a prompt generator to generate a prompt embedding from the input image. The prompt captures CNN's valuable inductive biases and fine-grained information, which guides AViT to achieve improved segmentation performance, particularly in scenarios with limited training data. By using ViT-Base as the ViT backbone, the number of tunable parameters of our AViT is 13.6M, which is only 13.7% of the total AViT's parameters.

Our contributions can be summarized as follows. (1) To the best of our knowledge, we are the first to introduce PEFT to directly mitigate ViTs' data-hunger in medical image segmentation. (2) We propose AViT, featuring adapters for transferring a pre-trained ViT to the SLS task and a prompt generator for enhancing information extraction. (3) The experimental results on 4 different public datasets indicate that AViT surpasses previous SOTA PEFT algorithms and ViT-based SLS models without pre-trained backbones (gains 2.91% and 2.32% on average IOU, respectively). Further, AViT achieves competitive, or even superior performance, to SOTA ViT-based SLS models with pre-trained backbones while having significantly fewer trainable parameters (13.6M vs. 143.5M).

2 Methodology

In skin lesion segmentation (SLS), the model is required to predict a segmentation map $\boldsymbol{Y} \in \{0,1\}^{H \times W}$ that partitions lesion areas based on an RGB skin image $\boldsymbol{X} \in \mathbb{R}^{H \times W \times 3}$. In Fig. 1-a, AViT applies a ViT backbone pre-trained on large natural image datasets to the downstream SLS task through adapters and a prompt generator and only optimizes a few parameters. We briefly describe the plain ViT backbone in Section 2.1 and discuss the details of AViT in Section 2.2.

2.1 Basic ViT

A plain ViT [10] backbone contains a patch embedding module and L transformer layers (Fig. 1-b). Given an image $\boldsymbol{X}$, the patch embedding module first splits the image into N non-overlapping patches, then flattens and maps them to D-dimensional patch embeddings $\boldsymbol{x} \in \mathbb{R}^{N \times D}$ through a linear projection, where $N = \frac{HW}{P^2}$ is the number of patches, and (P,P) is the patch size. The embedding sequence is then prepended with a learnable [`class`] token $\boldsymbol{x}_{class}$

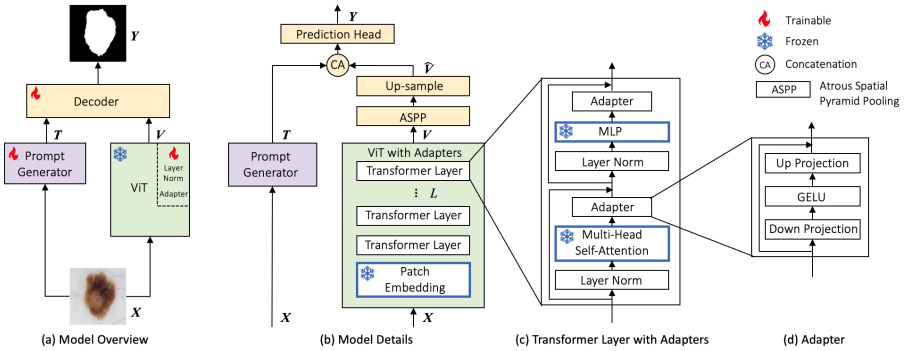

Fig. 1. Architecture of AViT: (a) Model overview with its prompt generator (a shallow CNN network), a large pre-trained ViT backbone with adapters, and a compact decoder. (b) Model details. (c) Details of a transformer layer with adapters. (d) Details of our adapters. During training, all modules in (b,c,d) contoured with blue borders are frozen, which encompasses 86.3% of AViT's parameters.

to get $\boldsymbol{x}_0 = [\boldsymbol{x}_{class}; \boldsymbol{x}] \in \mathbb{R}^{(N+1)\times D}$. To utilize the spatial prior, learnable position embeddings $\boldsymbol{E}_{pos} \in \mathbb{R}^{(N+1)\times D}$, defined in [10], are added to $\boldsymbol{x}_0$ to get $\boldsymbol{z}_0 = \boldsymbol{x}_0 + \boldsymbol{E}_{pos}$, which is the input of the first transformer layer. Each transformer layer (Fig. 1-c without the adapters) comprises a multi-head self-attention module (MSA) and a multi-layer perceptron module (MLP), along with layer norm (LN). The output of the lth transformer layer $\boldsymbol{z}_l \in \mathbb{R}^{(N+1)\times D}$ is:

$$\boldsymbol{z}'_l = MSA(LN(\boldsymbol{z}_{l-1})) + \boldsymbol{z}_{l-1} \tag{1}$$

$$\boldsymbol{z}_l = MLP(LN(\boldsymbol{z}'_l)) + \boldsymbol{z}'_l. \tag{2}$$

After getting the output of the final transformer layer $\boldsymbol{z}_L$, we remove its [`class`] token and reshape it to a 2D feature representation $\boldsymbol{V} \in \mathbb{R}^{\frac{H}{P}\times\frac{W}{P}\times D}$.

2.2 AViT

Given a pre-trained ViT backbone, we integrate adapters in each transformer layer to adjust the generated feature representation $\boldsymbol{V}$ adapted to skin images while leaving the weights of the backbone fixed. In addition, to enhance the information extraction, we employ a prompt generator in parallel, which is a shallow CNN network that produces a prompt embedding $\boldsymbol{T}$ based on the input image. Finally, a lightweight decoder combines $\boldsymbol{V}$ and $\boldsymbol{T}$ to predict a segmentation map. During training, we solely optimize the adapters, prompt generator, layer norm in the ViT backbone, and decoder, which collectively account for 13.7% of AViT's parameters. The details of these extensions are as follows.

Adapter Tuning: Similar to [20], we insert the adapter after MSA and MLP of each transformer layer (Fig. 1-c). The adapter (Fig. 1-d) contains two linear layers and a GELU function, which first projects the D-dimensional input into

a smaller dimension $\frac{D}{r}$, where r is the reduction ratio, and projects it back to D dimension, i.e., $Adapter(input) = GELU(input \cdot \boldsymbol{W}_{down})\boldsymbol{W}_{up}$. $\boldsymbol{W}_{down} \in \mathbb{R}^{D \times \frac{D}{r}}$ and $\boldsymbol{W}_{down} \in \mathbb{R}^{\frac{D}{r} \times D}$. The output of lth transformer layer with adapters is:

$$\boldsymbol{z}'_l = Adapter(MSA(LN(\boldsymbol{z}_{l-1}))) + \boldsymbol{z}_{l-1} \tag{3}$$

$$\boldsymbol{z}_l = Adapter(MLP(LN(\boldsymbol{z}'_l))) + \boldsymbol{z}'_l. \tag{4}$$

Information Enhancement by Prompt Tuning: Inspired by the prompt tuning [21,13], we deploy soft prompts to extract more information from images and enrich the SLS task learning. Specifically, we utilize the first stage of a ResNet-34 (including 7 convolutional layers) as the prompt generator to automatically create a prompt embedding $\boldsymbol{T}$ from the input image. The prompt is hypothesized to grasp CNN's helpful inductive biases and fine-grained information, e.g., spatial details, boundaries, and texture, to facilitate AViT's segmentation ability despite the small training datasets. Our soft prompt produced by the network is more flexible, customized to each input image, and includes rich information, in contrast to previous soft prompts that are simple free tunable parameters and remain constant for all inputs. Moreover, it is worth noting that our prompt generator has only a small number of parameters (0.23M), which is different from previous hybrid models combining a ViT with a large CNN backbone, e.g., ResNet-34 (21.3M) [37,19] or ResNet-50 (23.5M) [36].

Lightweight Decoder: We incorporate a compact decoder for efficient prediction, as opposed to prior works that use complex decoding architectures involving multi-stage up-sampling, convolutional operations, and skip connections [18,37]. This choice is driven by the powerful and over-parameterized nature of large pre-trained ViT backbones, which have demonstrated strong transferability to downstream tasks [28]. As visualized in Fig. 1-b, after getting the feature representation $\boldsymbol{V}$ from the ViT backbone and the prompt embedding $\boldsymbol{T}$ from the prompt generator, we first pass $\boldsymbol{V}$ through the atrous spatial pyramid pooling module (ASPP) proposed in [6], which uses multiple parallel dilated convolutional layers with different dilation rates, to obtain a feature that extracts local information while capturing lesion context at different scales. After that, we upsample the output feature of ASPP to get $\hat{\boldsymbol{V}}$, which has the same resolution as $\boldsymbol{T}$. Finally, $\hat{\boldsymbol{V}}$ is concatenated with $\boldsymbol{T}$ and sent to a projection head, which is formed by 3 convolutional layers connected by ReLU activation functions.

3 Experiments

Datasets and Evaluation Metrics: We evaluate our AViT on 4 public SLS databases collected from different sources: ISIC 2018 (ISIC) [8], Dermofit Image Library (DMF) [3], Skin Cancer Detection (SCD) [16], and PH2 [29], which contain 2594, 1212, 206, and 200 skin images along with their segmentation maps, respectively. We perform 5-fold cross-validation and measure our model's segmentation performance using Dice and IOU metrics, computational cost at inference via gigaFLOPs (GFLOPs), and memory footprint via the number of

Table 1. Skin lesion segmentation (SLS) results comparing BASE (AViT w/o both adapters and the prompt generator and is fully fine-tuned), AViT, and SOTA algorithms. We report the models' parameter count in millions (M). The 2nd column shows which pre-trained backbone the model used. R-34/50 represents ResNet-34/50.

Model	Pre-trained backbone	#Total Param. (M) ↓	#Tuned Param. (M) ↓	GFLOPs ↓	Segmentation Results in Test Sets (%)									
					Dice ↑					IOU ↑				
					ISIC	DMF	SCD	PH2	Avg$_{\pm std}$	ISIC	DMF	SCD	PH2	Avg$_{\pm std}$
(a) Full Fine-tuned BASE & Proposed PEFT Method														
BASE	ViT-B	91.8×	91.8×	18.0	90.77	91.69	91.95	95.64	92.51$_{0.38}$	83.71	84.89	85.42	91.72	86.43$_{0.34}$
AViT	ViT-B	99.4 (13.6×)	13.6×	20.9	91.74	92.04	93.16	95.66	93.15$_{0.42}$	85.22	85.47	87.39	91.72	87.45$_{0.70}$
(b) PEFT Methods														
VPT	ViT-B	92.8 (7.0×)	7.0×	26.5	90.89	91.26	89.09	93.14	91.10$_{0.46}$	83.83	84.14	80.76	87.27	84.00$_{0.74}$
AdaptFormer	ViT-B	93.0 (7.2×)	7.2×	18.2	91.12	91.27	89.65	93.76	91.45$_{0.42}$	84.15	84.18	81.49	88.33	84.54$_{0.67}$
(c) SLS Methods w/o Pre-trained Backbones & Trained From Scratch														
SwinUnet	None	41.4×	41.4×	8.7	89.64	90.67	89.77	94.24	91.08$_{0.57}$	81.94	83.19	82.07	89.24	84.11$_{0.79}$
UNETR	None	87.7×	87.7×	20.2	89.60	90.53	88.13	93.92	90.55$_{0.87}$	81.86	83.02	79.96	88.68	83.38$_{1.24}$
UTNet	None	10.0×	10.0×	13.2	89.68	89.87	88.11	93.29	90.23$_{0.61}$	81.99	81.91	79.71	87.62	82.81$_{0.77}$
MedFormer	None	19.2×	19.2×	13.0	90.47	90.85	90.60	94.82	91.68$_{0.74}$	83.22	83.52	83.53	90.23	85.13$_{1.12}$
Swin UNETR	None	25.1×	25.1×	14.3	90.19	91.00	90.71	94.54	91.61$_{0.49}$	82.78	83.77	83.54	89.74	84.96$_{0.74}$
(d) SLS Methods w/ Pre-trained Backbones & Fully Fine-tuned														
H2Former	R-34	33.7×	33.7×	24.7	91.17	91.29	92.76	95.65	92.72$_{0.63}$	84.35	84.22	87.04	91.77	86.85$_{0.91}$
FAT-Net	R-34, DeiT-T	28.8×	28.8×	42.8	91.26	91.32	93.03	96.07	92.92$_{0.48}$	84.42	84.25	87.23	92.48	87.10$_{0.80}$
BAT	R-50	46.2×	46.2×	10.3	91.33	91.20	92.95	95.84	92.83$_{0.46}$	84.40	84.03	87.08	92.04	86.89$_{0.78}$
TransFuse	R-50, DeiT-B	143.5×	143.5×	63.4	91.73	91.96	94.11	96.18	93.50$_{0.27}$	85.22	85.33	89.03	92.69	88.07$_{0.47}$

needed parameters. Due to the table width restriction and the high number of columns, we only report the standard deviation (std) of average Dice and IOU in tables and provide the std for each dataset in the supplementary material.

Implementation Details: We resize the images to 224×224 and augment them by random scaling, shifting, rotation, flipping, Gaussian noise, and brightness and contrast changes. The ViT backbone of AViT is a ViT-B/16 [10], with a patch size of 16×16, pre-trained on ImageNet-21k. Similar to [41], the reduction ratio r of the adapters is 4. The output dimension of ASPP is 256. All models are deployed on a single TITAN V and trained using a combination of Dice and binary cross entropy loss [11,33] for 200 epochs with the AdamW optimizer [26], a batch size of 16, 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 The Baseline (BASE): BASE is established by removing the adapters and the prompt generator of AViT and optimizing all the parameters during training. In Table 1-a, AViT achieves superior performance compared to BASE, with average IOU and Dice improvements of 1.02% and 0.64%, respectively, while utilizing significantly fewer trainable parameters (13.6M vs. 91.8M). This suggests that BASE exhibits overfitting, and full fine-tuning is unsuitable for transferring knowledge to smaller skin datasets, whereas AViT effectively leverages the learnt knowledge and demonstrates strong generalization capability on the SLS task. When considering the memory requirements for the 4 datasets, BASE would require storing 4 entirely new models, resulting in a total of $91.8\times4 = 367.2$M parameters. On the contrary, AViT only needs to store the pre-trained ViT backbone once, resulting in reduced storage needs, i.e., $85.8 + 13.6\times4 = 140.2$M. As the number of domains increases, the memory savings offered by AViT compared to BASE will become even more pronounced.

Comparing Against State-of-the-Art (SOTA) Methods: We conduct experiments on SOTA PEFT and SLS approaches. We first reproduced VPT [21]

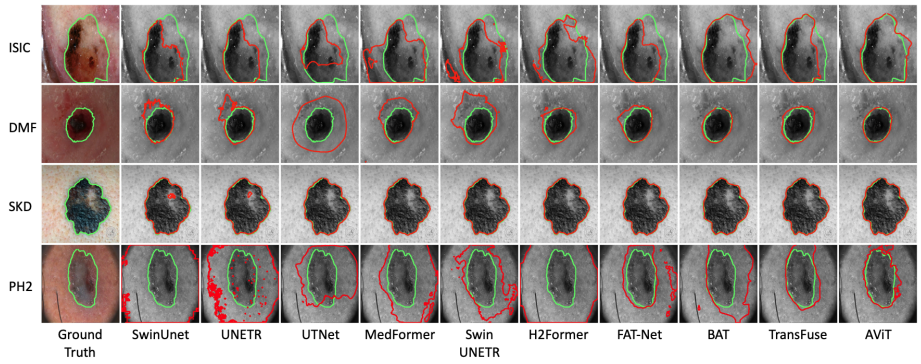

Fig. 2. Visual comparison with different SOTA methods. The green contours are the ground truth, and the red contours are the segmentation results.

that added learnable visual prompts in the input space and AdaptFormer [7] that introduced adapters in the transformer layers. We set the number of prompts in VPT to 100. In Table 1-b, AViT surpasses them across all datasets (gains 2.91% on average IOU over AdaptFormer), with comparable trainable parameters.

Additionally, we compare various ViT-based SLS algorithms and divide them into two groups. *Group 1* is models without pre-trained backbones and trained from scratch: SwinUnet [5], UNETR [18], UTNet [14], MedFormer [15], and Swin UNETR [34]. *Group 2* is models with pre-trained backbones and fully fine-tuned: H2Former [19], FAT-Net [37], BAT [36], and TransFuse [42]. H2Former and BAT used pre-trained ResNet but randomly initialized transformer modules. Table 1-c shows that AViT outperforms *Group 1* across all datasets by a large margin (increases average IOU of MedFormer by 2.32%), with comparable and even fewer trainable parameters (13.6M vs. 19.2M). Table 1-d illustrates that AViT achieves competitive or higher segmentation performance compared to *Group 2*, with fewer trainable parameters. For instance, AViT achieves a marginally lower average Dice compared to TransFuse (0.35% difference), yet its parameter count and computational complexity (GFLOPs) are 1/10 and 1/3 less than that of TransFuse, respectively. Fig. 2 visualizes AViT's segmentation performance.

AViT on Different Pre-trained ViT Backbones: We conduct experiments using ViTs in varied sizes, structures, or training strategies, including ViT-L/16, Swin-B, Swin-L [25], and DeiT-B, as the pre-trained backbone. For Swin-B/L, we use the output of its 3rd stage as the encoded image feature, whose resolution is the same as ViT-B's output feature. DeiT-B and ViT-B have the same architecture but different training strategies. In Table 2-a, for each ViT backbone, AViT achieves competitive and even higher performance compared to fully fine-tuned BASE, but with substantially fewer parameters (trainable and total) for the 4 datasets, indicating the applicability of our method on different ViTs.

Ablation Study: To show the efficacy of our proposed components in Section 2.2, we freeze the parameters of BASE's pre-trained ViT to get BASE* and

Table 2. Experiments using different pre-trained ViT backbones and ablation study of AViT. * means the pre-trained backbone is frozen throughout training. $^{-P}$ or $^{-A}$ represent not using the prompt generator or adapters in AViT.

Model	Pre-trained backbone	#Total Param. (M)	#Tuned Param. (M)	GFLOPs	Segmentation Results in Test Sets (%)									
					Dice ↑					IOU ↑				
					ISIC	DMF	SCD	PH2	Avg$_{\pm std}$	ISIC	DMF	SCD	PH2	Avg$_{\pm std}$
(a) Applicability to Various Pre-trained ViT Backbones														
BASE	Swin-B	63.8×	63.8×	15.6	91.63	91.70	92.71	95.88	92.98$_{0.37}$	85.05	84.89	86.60	92.13	87.17$_{0.61}$
AViT	Swin-B	68.9 (9.5×)	9.5×	18.3	91.54	91.73	93.60	95.68	93.14$_{0.39}$	84.90	84.94	88.12	91.77	87.43$_{0.64}$
BASE	Swin-L	139.8×	139.8×	32.3	91.64	91.69	92.93	95.83	93.02$_{0.25}$	85.08	84.86	86.97	92.04	87.24$_{0.43}$
AViT	Swin-L	151.1 (17.6×)	17.6×	36.3	91.56	91.91	93.74	96.07	93.32$_{0.31}$	84.93	85.24	88.38	92.47	87.76$_{0.50}$
BASE	ViT-L	311.2×	311.2×	61.2	91.37	91.76	93.23	95.86	93.06$_{0.29}$	84.60	84.99	87.52	92.09	87.30$_{0.47}$
AViT	ViT-L	336.9 (33.7×)	33.7×	67.7	91.54	91.77	93.48	95.73	93.13$_{0.48}$	84.88	85.01	87.94	91.85	87.42$_{0.79}$
BASE	DeiT-B	91.8×	91.8×	18.0	91.48	91.82	93.63	95.83	92.94$_{0.32}$	84.77	85.10	86.53	92.04	87.11$_{0.58}$
AViT	DeiT-B	99.4 (13.6×)	13.6×	20.9	91.70	91.85	93.67	95.97	93.30$_{0.31}$	85.14	85.17	88.22	92.30	87.71$_{0.51}$
(b) Ablation Study														
BASE*	ViT-B	91.8 (6.0×)	6.0×	18.0	87.18	89.23	86.24	90.17	88.20$_{0.46}$	77.92	80.81	76.27	82.30	79.33$_{0.65}$
AViT^{-P}	ViT-B	98.9 (13.2×)	13.2×	19.4	91.47	91.80	91.18	94.75	92.30$_{0.31}$	84.74	85.04	83.98	90.09	85.96$_{0.48}$
AViT^{-A}	ViT-B	92.3 (6.5×)	6.5×	19.5	90.87	91.00	89.09	93.87	91.21$_{0.88}$	83.78	83.72	81.18	88.53	84.30$_{1.19}$
AViT	ViT-B	99.4 (13.6×)	13.6×	20.9	91.74	92.04	93.16	95.66	93.15$_{0.42}$	85.22	85.47	87.39	91.72	87.45$_{0.70}$

remove the adapters and prompt generator in AViT to get AViT^{-A} and AViT^{-P}, respectively. In Table 2-b, BASE* attains average Dice and IOU of 88.20% and 79.33%, respectively. However, it still falls far behind fully fine-tuned BASE with 92.51% and 86.43% on average Dice and IOU, respectively. After adding adapters to BASE (AViT^{-P}), the average Dice and IOU increase by 4.10% and 6.63%, respectively; after adding a prompt generator to BASE (AViT^{-A}), the average Dice and IOU increase by 3.01% and 4.97%, respectively. Finally, AViT achieves the highest segmentation results and significantly outperforms BASE* (increases average Dice and IOU by 4.95% and 8.12%, respectively) with only 7.6M more trainable parameters. The above results reveal that our proposed mechanisms boost the segmentation performance, and a combination of both performs best.

4 Conclusion

We propose AViT, a new method to alleviate ViTs' data-hunger and apply it on small skin lesion segmentation (SLS) datasets by employing a pre-trained ViT backbone whilst keeping computation and storage memory costs very low via parameter-efficient fine-tuning (PEFT). Specifically, we integrate adapters into the transformer layers to modulate the backbone's image representation without updating its pre-trained weights and utilize a prompt generator to produce a prompt embedding, which captures CNNs' inductive biases and fine-grained information to guide AViT for segmenting skin images on limited data. Our experiments on 4 datasets illustrate that AViT outperforms other PEFT methods and achieves comparable or even superior performance to SOTA SLS approaches but with considerably fewer trainable and total parameters. Moreover, the experiments using different ViT backbones and an ablation study showcase the applicability of AViT and the effectiveness of AViT's components. Future work will focus on improving AViT's architecture so that it can achieve SOTA segmentation performance while retaining computation and memory efficiency.

References

1. Adegun, A., Viriri, S.: Deep learning techniques for skin lesion analysis and melanoma cancer detection: a survey of state-of-the-art. Artificial Intelligence Review **54**, 811–841 (2021)
2. Bahng, H., Jahanian, A., et al.: Visual prompting: Modifying pixel space to adapt pre-trained models. arXiv preprint arXiv:2203.17274 (2022)
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. Birkenfeld, J.S., Tucker-Schwartz, J.M., et al.: Computer-aided classification of suspicious pigmented lesions using wide-field images. Computer methods and programs in biomedicine **195**, 105631 (2020)
5. Cao, H., Wang, Y., et al.: SwinUnet: Unet-like pure transformer for medical image segmentation. In: ECCV 2022 Workshops. pp. 205–218. Springer (2023)
6. Chen, L.C., Papandreou, G., et al.: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
7. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., et al.: AdaptFormer: Adapting vision transformers for scalable visual recognition. In: NeurIPS 2022 (2022)
8. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., 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)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR 2009. pp. 248–255. Ieee (2009)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR 2020 (2020)
11. Du, S., Bayasi, N., Harmarneh, G., Garbi, R.: MDViT: Multi-domain vision transformer for small medical image segmentation datasets. arXiv preprint arXiv:2307.02100 (2023)
12. Du, S., Hers, B., Bayasi, N., et al.: FairDisCo: Fairer AI in dermatology via disentanglement contrastive learning. In: ECCVW 2022. pp. 185–202. Springer (2022)
13. Gao, Y., Shi, X., Zhu, Y., Wang, H., et al.: Visual prompt tuning for test-time domain adaptation. arXiv preprint arXiv:2210.04831 (2022)
14. Gao, Y., Zhou, M., Metaxas, D.N.: UTNet: a hybrid transformer architecture for medical image segmentation. In: MICCAI 2021. pp. 61–71. Springer (2021)
15. Gao, Y., et al.: A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark. arXiv preprint arXiv:2203.00131 (2022)
16. 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)
17. Gulzar, Y., Khan, S.A.: Skin lesion segmentation based on vision transformers and convolutional neural networks—a comparative study. Applied Sciences **12**(12), 5990 (2022)
18. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., et al.: UNETR: Transformers for 3D medical image segmentation. In: WACV 2022. pp. 574–584 (2022)
19. He, A., Wang, K., et al.: H2Former: An efficient hierarchical hybrid transformer for medical image segmentation. IEEE Transactions on Medical Imaging (2023)
20. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., et al.: Parameter-efficient transfer learning for NLP. In: ICML 2019. pp. 2790–2799. PMLR (2019)

21. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: ECCV 2022. pp. 709–727. Springer (2022)
22. Kinyanjui, N.M., Odonga, T., et al.: Fairness of classifiers across skin tones in dermatology. In: MICCAI 2020. pp. 320–329. Springer (2020)
23. Kirillov, A., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
24. Li, J., Chen, J., Tang, Y., Wang, C., Landman, B.A., Zhou, S.K.: Transforming medical imaging with transformers? a comparative review of key properties, current progresses, and future perspectives. Medical image analysis p. 102762 (2023)
25. Liu, Z., Lin, Y., Cao, Y., Hu, H., et al.: Swin Transformer: Hierarchical vision transformer using shifted windows. In: ICCV 2021. pp. 10012–10022 (2021)
26. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
27. Maron, R.C., Hekler, A., Krieghoff-Henning, E., Schmitt, M., et al.: Reducing the impact of confounding factors on skin cancer classification via image segmentation: technical model study. Journal of Medical Internet Research **23**(3), e21695 (2021)
28. Matsoukas, C., Haslum, J.F., et al.: What makes transfer learning work for medical images: feature reuse & other factors. In: CVPR 2022. pp. 9225–9234 (2022)
29. Mendonça, T., Ferreira, P.M., et al.: PH 2-A dermoscopic image database for research and benchmarking. In: EMBC 2013. pp. 5437–5440. IEEE (2013)
30. Mirikharaji, Z., Abhishek, K., Bissoto, A., Barata, C., et al.: A survey on deep learning for skin lesion segmentation. Medical Image Analysis **88**, 102863 (2023)
31. Peters, M.E., Ruder, S., Smith, N.A.: To tune or not to tune? adapting pretrained representations to diverse tasks. ACL 2019 p. 7 (2019)
32. Siegel, R.L., Miller, K.D., Wagle, N.S., Jemal, A.: Cancer statistics, 2023. CA: a cancer journal for clinicians **73**(1), 17–48 (2023)
33. 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)
34. 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)
35. 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)
36. Wang, J., Wei, L., Wang, L., et al.: Boundary-aware transformers for skin lesion segmentation. In: MICCAI 2021. pp. 206–216. Springer (2021)
37. Wu, H., Chen, S., et al.: FAT-Net: Feature adaptive transformers for automated skin lesion segmentation. Medical image analysis **76**, 102327 (2022)
38. Wu, J., Fu, R., Fang, H., Liu, Y., Wang, Z., Xu, Y., Jin, Y., Arbel, T.: Medical SAM adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 (2023)
39. Xie, Y., Zhang, J., Xia, Y., Shen, C.: A mutual bootstrapping model for automated skin lesion segmentation and classification. IEEE transactions on medical imaging **39**(7), 2482–2493 (2020)
40. Yan, Y., Kawahara, J., Hamarneh, G.: Melanoma recognition via visual attention. In: IPMI 2019. pp. 793–804. Springer (2019)
41. Yang, T., Zhu, Y., Xie, Y., Zhang, A., Chen, C., Li, M.: AIM: Adapting image models for efficient video action recognition. In: ICLR 2023 (2023)
42. Zhang, Y., Liu, H., Hu, Q.: TransFuse: Fusing transformers and cnns for medical image segmentation. In: MICCAI 2021. pp. 14–24. Springer (2021)