Leveraging Inductive Bias in ViT for Medical Image Diagnosis

Jungmin Ha[†1]

Euihyun Yoon[†1]

Sungsik Kim[1]

Jinkyu Kim[*2]
jinkyukim@korea.ac.kr

Jaekoo Lee[*1]
jaekoo@kookmin.ac.kr

[1] College of Computer Science
Kookmin University
Seoul, Korea

[2] Department of Computer Science and Engineering
Korea University,
Seoul, Korea

Abstract

Recent advances in attention-based models have raised expectations for an automated diagnosis application in computer vision due to their high performance. However, attention-based models tend to lack some of the inherent assumptions for images, known as inductive biases, which convoultional-based models possess. Herein, we customize a vision transformer (ViT) model to enhance the performance with exploiting locality inductive biases for limited medical images. Specifically, using the ViT model as a backbone, we propose shift window attention (SWA), deformable attention (DA), and a convolutional block attention module (CBAM) to leverage the convolutional neural networks' inductive bias towards locality, thereby improving both global and local context of the lesion. To evaluate the effectiveness and efficiency of our proposed method, we use various publicly available well-known medical images diagnosis such as HAM10000, MURA, ISIC 2018 and CVC-Clinic DB for classification or dense prediction tasks. Experimental results show that our method significantly outperforms the other state-of-the-art alternatives. Furthermore, we utilize GradGAM++ to qualitatively visualize the image regions where the network attends to. Our code is available at Medical_CBAM_ViT.

1 Introduction

In the medical field, research on utilizing artificial intelligence (AI) for medical image diagnosis has actively been pursued [35]. The use of AI in medicine and healthcare offers the advantage of enhancing diagnostic accuracy and expediting the decision-making process for medical practitioners. Recent medical imaging encompasses a variety of visual modalities such as X-ray, magnetic resonance imaging (MRI), and computed tomography (CT).

[†] Equal Contribution.

[*] Corresponding Author.

Consequently, the integration of AI in medicine helps reduce diagnostic inaccuracies by medical specialists and streamlines diagnostic procedures, potentially leading to improved cost-effectiveness [13, 17, 19].

To date, in machine learning of medical images, convolutional neural network (CNN) architectures have predominantly been used for disease diagnosis [30, 33, 42] primarily due to relatively good generalization performance even with limited data, which is a crucial advantage given the lack of data in the medical field. To extract image features, a CNN applies convolutional filters that slide over the image, using locality and stationarity assumptions. Thus, CNNs are prone to performance degradation when objects are occluded or do not satisfy the locality conditions [4]. In contrast, in general computer vision, recent research has been focused on attention-based models through vision transformer (ViT) models [12, 34]. The ViT model performs attention operations over the entire image to capture the global context, leading to improved performance [8]. However, a drawback is that ViT models require more training time and data compared to CNNs, that exploit locality and stationarity, primarily due to the insufficient inductive bias of ViT models [21, 39].

Therefore, we propose a transformer-based architecture that leverages inductive bias for a variety of visual modalities in medical diagnosis. The proposed architecture incorporates a window-based attention operation [21] to add locality, which is an inductive bias inherent to images, while applying deformable attention operation [39] to capture more globally informative features. Additionally, a convolutional block attention module [37] is used at the network bottleneck of the proposed architecture to integrate both channel-wise and spatial-wise information of the features.

As a result, our method shows better performance against alternative state-of-the-art (SOTA) ViT models. To quantitatively evaluate our method, we first use publicly available medical imaging datasets such as HAM10000$_C$ [32], and MURA [25] for classification task as well as HAM10000$_S$ [32] and ISIC 2018 [7] for dense prediction task. We also analyze our method in detail to support our claims by using the qualitative visualization.

Our contributions are as follows: (i) In medical limited datasets, our method facilitates the inductive bias of ViT models by integrating shift window attention (SWA), deformable attention (DA), and a convolutional block attention module (CBAM). (ii) Our experiments show that our method quantitatively and qualitatively outperforms prior medical diagnosis methods in various tasks and datasets within a fair and realistic setting.

2 Related Work

2.1 Medical Image Diagnosis

In the early stages of automated medical image diagnosis, computer image processing techniques were employed with hand-designed feature extractors. Specifically, handcrafted features such as lesion color, shape, and size were used to discriminate various types of lesions [2, 11]. However, traditional learning methods based on feature engineering can introduce issues due to human subjectivity. To address these issues, there has been a gradual shift towards representation learning or deep learning [16, 20, 36].

Initial deep learning-based medical image diagnosis approaches primarily utilized CNN-based architectures structurally specialized for image data, relying on assumptions of locality and stationarity, yielding significant results. To leverage the semantic information of comprehensive images alongside local information, modified CNN-based methods specialized for medical imaging data have emerged [16, 20, 36]. These include DCNN [36], with kernel size variations tailored to medical imaging; GCCN [16], based on Gabor wavelet inner products; and FixCaps [20], which utilizes capsule networks to jointly capture channel and spatial information, achieving better performance in dense prediction tasks. However, these CNN-based methods still face limitations in effectively capturing global semantic information [23].

Recently, ViT models have gained prominence by learning the global context of images via attention mechanisms, enabling them to extract more comprehensive semantic information. In medical computer vision, where precise diagnosis requires leveraging both local and global information, approaches using ViT models have been emerging[1, 10, 22].

2.2 Vision Transformer-based Architectures

The transformer architecture was first introduced in natural language processing and exhibited overwhelming performance across all language tasks[34]. In the field of computer vision, ViT models emerged, and demonstrated overwhelming performance across various vision tasks [8]. ViT models divide a image into patches and apply attention between them, offering advantages in learning global context. However, compared to CNNs, it has limitations due to a lack of locality inductive bias.

To address the issue of insufficient locality inductive bias in ViT models, of a landmark work, the Swin Transformer [21] model employs a hierarchical architecture and shifted window attention, while DeiT [31] uses knowledge distillation to leverage the knowledge of pre-trained CNNs. CvT [38] is a hybrid model that adds convolutional blocks to the transformer architecture. Similarly, in the medical computer vision domain, hybrid CNN and ViT models like MedViT [22] have achieved SOTA performance.

3 Method

We propose a enhanced ViT model that leverages locality inductive bias to comprehensively understand not only the local lesion areas in a medical image but also the relationships between global regions surrounding the lesion. Our model is particularly effective in medical computer vision, where understanding the overall organ appearance as well as the local lesion areas is important for a more accurate diagnosis.

Our model utilizes the following three modules to address locality inductive bias: (i) stem block, (ii) shift window attention (SWA) block, and (iii) deformable attention (DA) block. Our stem block consists of two convolutional layers to preprocess the input image, capturing low-level features (Section 3.1). SWA block uses attention-based layers to extract local information, guiding a model toward having locality inductive bias (Section 3.2). DA block uses deformable attention layers to extract spatial cues spanning a larger area (Section 3.3). Moreover, to enable the extraction of diverse feature maps, convolutional block attention module (CBAM) is incorporated at the last of every stage (Section 3.4). Our architecture is shown in Figure 1.

3.1 Stem Block

In the original ViT model, an input image is split into fixed-size patches, each linearly embedded. Sequence of the embedded patches is then fed into a transformer encoder. Patch-wise

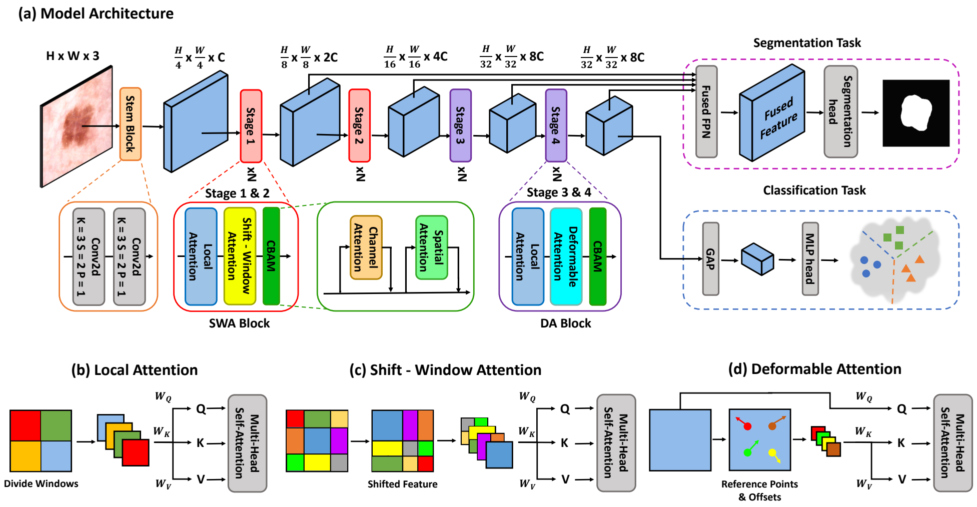

Figure 1: (a) An overview of our proposed model. Built upon Vision Transformer, we use the following three building blocks: (1) Stem Block, (2) SWA Block for 1st and 2nd stages, and (3) DA Block for 3rd and 4th stages. In image classification, the output feature map undergoes Global Average Pooling(GAP) and MLP processing. For segmentation, fused feature maps with Fused Feature Pyramid Network(FPN) from Stages are utilized. (b, c, d) Detailed Explanation of Local Attenton, Shifted-Window Attention and Deformable Attention

encoding has the advantage of learning rich representations. Similar to the prevalent practice in numerous prior ViT models, we employed two 3×3 convolution layers as a preprocessor to encode low-level features [41].

3.2 Shift Window Attention (SWA) Block

Common ViT models, which struggle to leverage local inductive bias, generally require significantly huge training datasets compared to CNNs to achieve comparable performance. In healthcare and medical domain, data acquisition is often challenging due to strict regulatory requirements and the need for knowledge-intensive guidance from medical specialists. Therefore, a modified architecture that can leverage inductive bias is required to effectively apply the ViT model to medical image diagnosis.

We propose to use the shift window attention (SWA) block, which utilizes a localized attention module to induce the effect of locality inductive bias. Supposing each window has $M \times M$ size, the computational complexity of a SWA is computed as

$$\Omega(\mathrm{SWA}) = 4HWC^2 + 2M^2HWC \tag{1}$$

where C is the feature dimension and input image size is $H \times W$.

As shown in Figure 1, our SWA block consists of three main layers: (i) local attention, (ii) shift window attention, and (iii) CBAM.

Local Attention: As shown in Figure 1 (b), our local attention layer uses a window of pre-determined size to divide the feature map, generating attention weights accordingly.

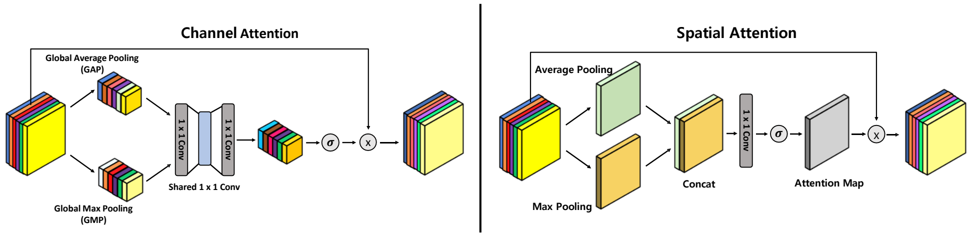

Figure 2: Detailed architecture of channel and spatial attentions.

And self-attention has been seperately applied on each divided windows, similar to what is depicted. To incorporate relative positional information for each window, we add relative positional encoding to divided windows while calculating self-attention. Thus, we constrain the model to extract local features, injecting appropriate locality inductive bias.

Shift-Window Attention: As shown in Figure 1 (c), our shift window attention layer then operates between local features, generating attention between windows and allowing the model flexibility in its representation. During this process, the feature map is divided as shown in Figure 1 (c), then the windows are rotated similar to what is depicted. And following the same process as in Figure 1 (b), self-attention is applied.

3.3 Deformable Attention (DA) Block

The previous SWA block focuses on extracting local information by using a window-based constraint, but a model also needs to understand global (or contextual) information to recognize legions, which often span broad areas. To effectively extract such contextual information, we use a deformable attention layer, which can flexibly see wider regions with a larger receptive field. Similar to the previous SWA block, our DA block consists of three layers: (i) local attention, (ii) deformable attention, and (iii) CBAM layer. We use the same layers for (i) and (iii) with the SWA block. The deformable attention layer is highly influenced by the size of the feature map since it samples reference points from the feature map and then computes offsets through a sub-network. Therefore, it is applied towards the latter part of the model where the size of the feature map becomes relatively smaller.

Deformable Attention: As shown in Figure 1 (d), our deformable attention layer uniformly samples reference points from the feature map, and then uses offsets from the sampled points to flexibly select features. Offsets are computed by an sub-offset network which composed of two CNN layers. The values corresponding to the selected points through the offsets are used as the key and value in the multi-head attention. And a linear operation is applied to the entire feature map to use as query. Through this process, our model can focus on relevant areas with a flexible window, capturing contextual and global information. The computational complexity of DA is

$$\Omega(\mathrm{DA}) = 2HWN_sC + 2HWC^2 + 2N_sC^2 + (k^2+2)N_sC \tag{2}$$

where N_s is the number of sampled points and k denotes the kernel size of convolution layers in the sub-network that calculates the offset.

Dataset	HAM10000$_C$ [32]	HAM10000$_S$ [32]	MURA [25]	ISIC 2018 [7]	CVC-Clinic DB [5]
Train/Val./Test	51,646 / 1,006 / 828	9,187 / - / 828	36,608 / - / 3,197	2,594 / - / 100	551 / - / 61
Task	Classification	Segmentation	Classification	Segmentation	Segmentation
Type of diagnosis	Pigmented skin lesions	Pigmented skin lesions	Fracture	Dermatoscopy	Colon polyp

Table 1: Description of datasets used in experiments.

Type	Networks	Details			Classification		Segmentation			
		Resolution	Params(MACs)	FPS	HAM10000$_C$	MURA	+Decoder	HAM10000$_S$	ISIC 2018	CVC-Clinic
CNN	ResNet50[14]	224×224	25.6M(4.1G)	184.0	0.973 (0.977)	0.720	UperNet	0.925 (0.919)	0.882(0.886)	0.849
	GoogLeNet[28]	224×224	13.0M(1.5G)	152.2	0.940 (0.964)	0.749	UperNet	0.925 (0.922)	0.879 **(0.889)**	0.839
	Inception V3[29]	299×299	27.2M(2.9G)	100.6	0.973 (0.952)	0.713	UperNet	0.922 (0.921)	0.879 (0.888)	0.838
	MobileNet V3[15]	224×224	5.5M(0.2G)	171.1	0.971 (0.973)	0.710	UperNet	0.924 (0.919)	0.897 (0.894)	0.820
	FixCaps[20]	299×299	0.8M(1.4G)	81.5	0.961 (0.966)	0.683	Autoencoder	0.747 (0.739)	0.737 (0.738)	0.752
ViT	ViT-B/32[8]	224×224	88.2M(4.4G)	146.3	0.960 (0.955)	0.654	UperNet	0.908 (0.905)	0.871 (0.866)	0.770
	Swin-B[21]	224×224	87.8M(10.2G)	27.0	0.954 (0.936)	0.743	UperNet	0.928 (0.924)	0.884 (0.880)	0.813
	MedViT[22]	224×224	45.4M(8.44G)	36.4	0.930 (0.907)	**0.787**	UperNet	0.838 (0.842)	0.825 (0.857)	**0.874**
	Ours	224×224	59.7M(15.8G)	32.0	**0.982 (0.978)**	0.754	UperNet	**0.940 (0.933)**	**0.898** (0.888)	0.854

Table 2: Comparison of classification and segmentation performance on various datasets. Note that scores in parenthesis represent results with the black-hat transform as preprocessing. Bold text indicates the best performance, while underlined text indicates the second-best performance among all models.

3.4 Convolutional Block Attention Module (CBAM)

In a transformer architecture, attention is conducted on the spatial dimensions $H \times W$ from an input image $I \in \mathbb{R}^{H \times W \times C}$. Therefore, the operations across the spatial dimensions lead to a diminished correlation between features extracted across the channels. This is particularly crucial in tasks like segmentation, where the channel dimension detects what object is present, and each channel's $H \times W$ dimensions detect where the object is located [24, 37]. Hence, in this paper, we design the modified transformer architecture with applied inductive bias to apply CBAM at each bottleneck. This adaptive utilization of the extracted features enables the model to achieve high performance not only in classification but also in dense prediction tasks such as segmentation [9].

As shown in Figure 2, our CBAM layer consists of channel-wise and spatial-wise attention with a skip connection. Therefore, by integrating these two types of information, each with its distinct features, it can effectively extract the refined representations. We employ channel attention and spatial attention separately, using the CBAM module for input feature maps with negligible overheads. In particular, given an input $I \in \mathbb{R}^{H \times W \times C}$, each attention operation calculates the average and maximum values, at both the spatial and channel levels, as shown in Figure 2.

4 Experiments

4.1 Datasets

We use the following datasets for the performance evaluation of the model in various publicly available datasets : HAM10000$_C$, HAM10000$_S$ [32], MURA [25], ISIC 2018 [7], and CVC-Clinic DB [5]. We provide about statistics and details of used datasets in Table 1.

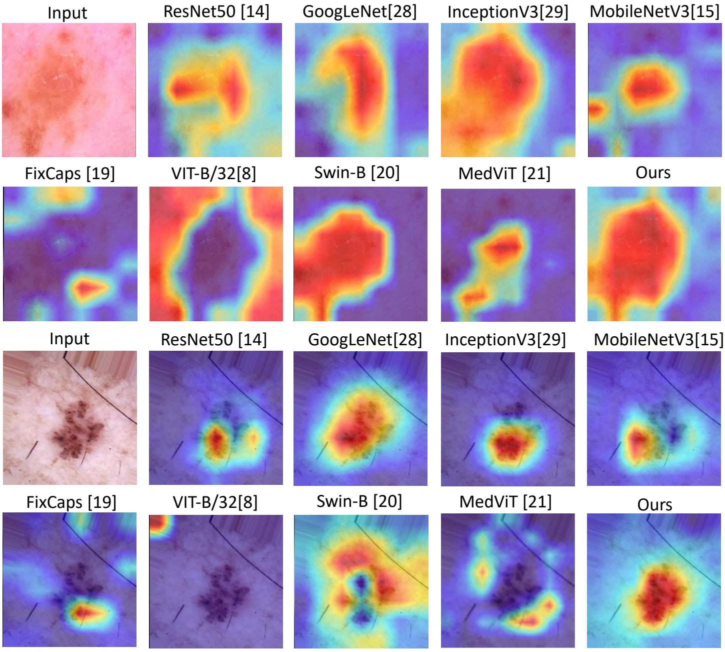

Figure 3: Visualization comparison using GradCAM++ [6] for ours and alternative models. *Data*: HAM10000$_S$

4.2 Experimental Results

Classification. Based on HAM10000$_C$ [32] and MURA [25] datasets, we train our model end-to-end from scratch, and we compare its classification top-1 accuracy with other alternative models such as CNN-based and ViT-based models.

We observe in Table 2 for the HAM10000$_C$ [32] dataset that our proposed model generally outperforms the SOTA visual recognition models in terms of average top-1 accuracy with a large gain. Ours shows 98.2% in accuracy, which is 0.9-5.2% higher than alternatives. For the MURA [25] dataset, our model shows the second-best performance to the SOTA model while outperforming other CNN-based models. In particular, ViT-B/32, when trained on limited medical imaging datasets, demonstrates relatively lower performance compared to CNN-based models due to its lack of inducitve bias.

Dense Prediction. To evaluate the performance in dense prediction for each model, we also conduct a segmentation task. We use mean intersection over union (mIoU) metric to evaluate the segmentation task. Each backbone model, pre-trained on image classification data, is fine-tuned for segmentation by connecting it to a segmentation decoder and training the entire model. We use a UperNet [40] which is Unet [26] based decoder. In the UperNet [40] architecture, segmentation is performed by utilizing four different feature maps of different sizes obtained from the backbone network. As depicted in Figure 1, we extract feature maps from the corresponding stage of each backbone network and employ them as inputs for the segmentation process of UperNet [40]. Note that FixCaps [20] is built upon CapsNets [27], which is not possible to use UperNet-style multi-scale decoders. Instead, we use the standard decoder architecture from Autoencoder [18].

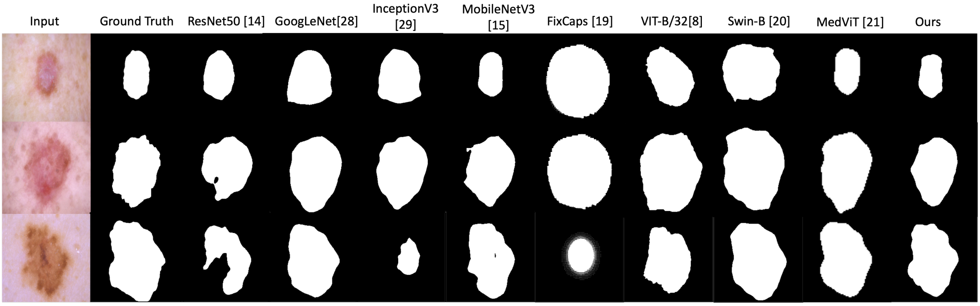

Figure 4: Visualization of ground truth and corresponding segmentation results obtained by ours and other alternative models. *Data*: HAM10000$_S$

As shown in Table 2, we present the evaluation results of segmentation performance on the HAM10000$_S$ [32] dataset, ISIC 2018 [7] and CVC-Clinic DB [13] using a backbone model pre-trained on HAM10000$_C$ [32]. Looking at the performance on both HAM10000$_S$ [32] and ISIC 2018 [7], ours outperforms other models. Notably, ours shows better performance than the Swin-B [21], which uses shift-window attention to alleviate inductive bias of the ViT model. This indicates that the SWA and DA blocks, along with CBAM in our model, capture not only ViT's global context but also local context through channel and spatial attention, making it effective for dense prediction [9].

To assess the performance of the model in various datasets, we utilize colonoscopy data from CVC-Clinic DB [13]. The backbone model is pre-trained on HAM10000$_C$ [32]. This decision was made because diagnosing skin cancer and colonoscopy both require color and shape information under similar conditions. As shown in Table 2 for the CVC-Clinic DB [13], ours demonstrates the comparable performance to other SOTA model on colonoscopy data. Particularly noteworthy is the performance difference with ViT-B/32, which does not add inductive bias, showing an 8.4% gap, and with Swin-B [21], showing a 4.1% difference. This interpretation can be attributed to the effective extraction of feature maps, which are more suitable for dense prediction using CBAM [9].

We conducted additional experiments with the HAM10000$_C$, HAM10000$_S$ [32], ISIC 2018 [7] datasets using black-hat transform preprocessing. The black-hat transform is a preprocessing technique that removes irrelevant elements of the lesion such as hair, which could affect dermatoscopy diagnosis [3]. The parentheses in Table 2 show a comparison of all models applying the black-hat transformation preprocessing.

For CNN-based models, the performance of those applying the black-hat transform preprocessing increases compared to those that do not apply it in most cases. In contrast, for ViT-based models, applying the black-hat transform preprocessing does not enhance performance. These results suggest that removing irrelevant elements helps only CNN-based models by allowing them to focus on the lesion. However, our method guarantees the best performance regardless of whether the black-hat transform preprocessing is applied.

As shown in Figure 3, we apply GradCAM++ [6] to qualitatively analyze perceptual regions, which the model focuses on for the final verdict. The well-known GradCAM++ [6] is a visualization technique that utilizes gradients from the output to determine which regions of the image are deemed important by the network. We observe that our model demonstrates a heightened concentration on clinically significant regions associated with lesions compared to other models.

From the visualizations of GradCAM++ [6] in Figure 3, we observe that while CNN-based models generally attend to regions near the lesions, ViT-based models such as ViT-B/32, Swin-B [21], and MedViT [22] fail to focus on the lesion area. This indicates that existing ViT-base models, which lack locality inductive bias, are skewed toward less important global information. However, our model, compared to both CNN and ViT models, effectively distinguishes important regions of the lesions by appropriately utilizing both local and global contexts.

As shown in Figure 4, which visualizes the segmentation results on the HAM10000$_S$ dataset, our model produces segmentation results that are closest to the ground truth.

5 Conclusion

In this paper, we utilized inductive bias to apply a ViT model with high generalization performance in the medical computer vision domain, where data is limited. To enhance performance in the classification and dense prediction tasks, which involves understanding both the global and local context of lesions in medical diagnosis, we modified the ViT model using SWA, DA, and CBAM.

As a result, our proposed method shows comparable performance to other SOTA alternatives such as CNN-based and ViT-based models. The performance comparable to that of the SOTA visual recognition models confirmed that leveraging inductive bias in ViT is effective for medical image diagnosis. Furthermore, to validate stability, visualization technique was employed to verify whether the predictions made by each model were based on valid regions in the image. The results confirmed that ours made diagnoses based on valid areas compared to other comparing models, providing evidence of the model's stability.

Our work significantly contributes to enhancing diagnostic accuracy not only in the field of skin cancer but also in various other medical computer vision domains, such CT and MRI. Furthermore, the potential expansion of our approach holds promising prospects for next-generation of healthcare and medicine through large language models and large multi-modal models based on the transformer architecture.

Acknowledgement. This research was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant (No.RS-2022-00167194; Trustworthy AI on Mission Critical Systems, IITP-2024-RS-2024-00357879; AI-based Biosignal Fusion and Generation Technology for Intelligent Personalized Chronic Disease Management) funded by the Korea government(MSIT). This work was also supported by the Mid-career Researcher Program through the National Research Foundation of Korea(NRF), funded by the Ministry of Science and ICT (No.RS-2024-00352326).

References

[1] Saad Aladhadh et al. An effective skin cancer classification mechanism via medical vision transformer. Sensors, 22(11):4008, 2022.

[2] Samuel G Armato III et al. Automated detection of lung nodules in ct scans: preliminary results. Medical physics, 28(8):1552–1561, 2001.

[3] Hritam Basak et al. Mfsnet: A multi focus segmentation network for skin lesion segmentation. Pattern Recognition, 128:108673, 2022.

[4] Peter W Battaglia et al. Relational inductive biases, deep learning, and graph networks. 2018.

[5] Jorge Bernal et al. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized medical imaging and graphics, 43:99–111, 2015.

[6] Aditya Chattopadhay et al. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 839–847, 2018. doi: 10.1109/WACV.2018.00097.

[7] Noel Codella 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.

[8] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[9] Jun Fu et al. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3146–3154, 2019.

[10] Yunhe Gao et al. Utnet: a hybrid transformer architecture for medical image segmentation. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24, pages 61–71. Springer, 2021.

[11] A Green et al. Computer image analysis of pigmented skin lesions. Melanoma research, 1(4):231–236, 1991.

[12] Kai Han et al. A survey on vision transformer. IEEE transactions on pattern analysis and machine intelligence, 45(1):87–110, 2022.

[13] Palak Handa et al. Automatic detection of colorectal polyps with mixed convolutions and its occlusion testing. Neural Computing and Applications, 35(26):19409–19426, 2023.

[14] Kaiming He et al. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[15] Andrew Howard et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF international conference on computer vision, pages 1314–1324, 2019.

[16] Qian Huang, Wei Li, et al. Blood cell classification based on hyperspectral imaging with modulated gabor and cnn. IEEE Journal of Biomedical and Health Informatics, 24(1):160–170, 2019.

[17] Kumari Jyoti et al. Automatic diagnosis of covid-19 with mca-inspired tqwt-based classification of chest x-ray images. Computers in Biology and Medicine, 152:106331, 2023.

[18] Sadegh Karimpouli and Pejman Tahmasebi. Segmentation of digital rock images using deep convolutional autoencoder networks. volume 126, pages 142–150. Elsevier, 2019.

[19] Inkyung Kim et al. Fall detection using biometric information based on multi-horizon forecasting. In 2022 26th International Conference on Pattern Recognition (ICPR), pages 1364–1370. IEEE, 2022.

[20] Zhangli Lan et al. Fixcaps: An improved capsules network for diagnosis of skin cancer. IEEE Access, 10:76261–76267, 2022.

[21] Ze Liu et al. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision, pages 10012–10022, 2021.

[22] Omid Nejati Manzari et al. Medvit: a robust vision transformer for generalized medical image classification. Computers in Biology and Medicine, 157:106791, 2023.

[23] Christos Matsoukas et al. Is it time to replace cnns with transformers for medical images? arXiv preprint arXiv:2108.09038, 2021.

[24] Lei Mou et al. Cs-net: Channel and spatial attention network for curvilinear structure segmentation. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22, pages 721–730. Springer, 2019.

[25] Pranav Rajpurkar et al. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. arXiv preprint arXiv:1712.06957, 2017.

[26] Olaf Ronneberger et al. U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pages 234–241. Springer, 2015.

[27] Sara Sabour et al. Dynamic routing between capsules. Advances in neural information processing systems, 30, 2017.

[28] Christian Szegedy et al. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1–9, 2015.

[29] Christian Szegedy et al. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2818–2826, 2016.

[30] Nikhil Kumar Tomar et al. Fanet: A feedback attention network for improved biomedical image segmentation. IEEE Transactions on Neural Networks and Learning Systems, 2022.

[31] Hugo Touvron et al. Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning, pages 10347–10357. PMLR, 2021.

[32] Philipp Tschandl et al. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data, 5(1):1–9, 2018.

[33] Jeya Maria Jose Valanarasu and Vishal M Patel. Unext: Mlp-based rapid medical image segmentation network. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V, pages 23–33. Springer, 2022.

[34] Ashish Vaswani et al. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[35] Xuesong Wang et al. Contrastive functional connectivity graph learning for population-based fmri classification. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I, pages 221–230. Springer, 2022.

[36] Y. Wang et al. An optimized deep convolutional neural network for dendrobium classification based on electronic nose. Sensors and Actuators A: Physical, 307:111874, 2020.

[37] Sanghyun Woo et al. Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), pages 3–19, 2018.

[38] Haiping Wu et al. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 22–31, 2021.

[39] Zhuofan Xia et al. Vision transformer with deformable attention. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 4794–4803, 2022.

[40] Tete Xiao, Yingcheng Liu, et al. Unified perceptual parsing for scene understanding. In Proceedings of the European conference on computer vision (ECCV), pages 418–434, 2018.

[41] Tete Xiao et al. Early convolutions help transformers see better. Advances in Neural Information Processing Systems, 34:30392–30400, 2021.

[42] Qing Xu et al. Dcsau-net: A deeper and more compact split-attention u-net for medical image segmentation. Computers in Biology and Medicine, 154:106626, 2023.