A Novel Multi-task Model Imitating Dermatologists for Accurate Differential Diagnosis of Skin Diseases in Clinical Images

Yan-Jie Zhou[1,2](✉), Wei Liu[1,2], Yuan Gao[1,2], Jing Xu[1,2], Le Lu[1], Yuping Duan[3], Hao Cheng[4], Na Jin[4], Xiaoyong Man[5], Shuang Zhao[6], and Yu Wang[1](✉)

[1] DAMO Academy, Alibaba Group, Hangzhou, China
zhouyanjie.zyj@alibaba-inc.com, Flimanadam@gmail.com
[2] Hupan Lab, Hangzhou, China
[3] School of Mathematical Sciences, Beijing Normal University, Beijing, China
[4] Sir Run Run Shaw Hospital, Hangzhou, China
[5] The Second Affiliated Hospital Zhejiang University School of Medicine, Hangzhou, China
[6] Xiangya Hospital Central South University, Changsha, China

Abstract. Skin diseases are among the most prevalent health issues, and accurate computer-aided diagnosis methods are of importance for both dermatologists and patients. However, most of the existing methods overlook the essential domain knowledge required for skin disease diagnosis. A novel multi-task model, namely **DermImitFormer**, is proposed to fill this gap by imitating dermatologists' diagnostic procedures and strategies. Through multi-task learning, the model simultaneously predicts body parts and lesion attributes in addition to the disease itself, enhancing diagnosis accuracy and improving diagnosis interpretability. The designed lesion selection module mimics dermatologists' zoom-in action, effectively highlighting the local lesion features from noisy backgrounds. Additionally, the presented cross-interaction module explicitly models the complicated diagnostic reasoning between body parts, lesion attributes, and diseases. To provide a more robust evaluation of the proposed method, a large-scale clinical image dataset of skin diseases with significantly more cases than existing datasets has been established. Extensive experiments on three different datasets consistently demonstrate the state-of-the-art recognition performance of the proposed approach.

Keywords: Skin disease · Multi-task learning · Vision transformer

1 Introduction

As the largest organ in the human body, the skin is an important barrier protecting the internal organs and tissues from harmful external substances, such as

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14225, pp. 202–212, 2023.
https://doi.org/10.1007/978-3-031-43987-2_20

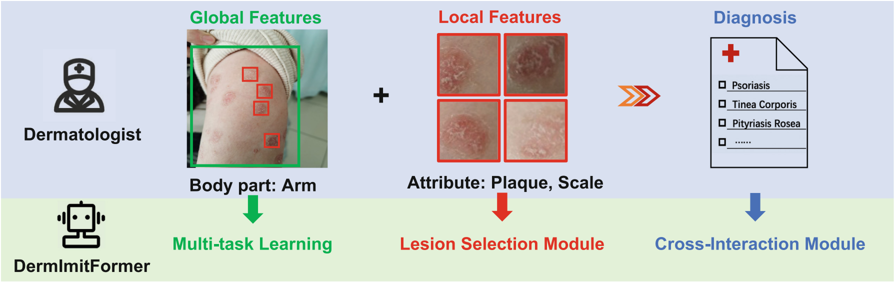


Fig. 1. The relationship between dermatologists' diagnostic procedures and our proposed model (best viewed in color).

sun exposure, pollution, and microorganisms [8,10]. In recent years, the increasing number of deaths by skin diseases has aroused widespread public concern [16,17]. Due to the complexity of skin diseases and the shortage of dermatological expertise resources, developing an automatic and accurate skin disease diagnosis framework is of great necessity.

Among non-invasive skin imaging techniques, dermoscopy is currently widely used in the diagnosis of many skin diseases [1,7], but it is technically demanding and not necessary for many common skin diseases. Clinical images, on the contrary, can be easily acquired through consumer-grade cameras, increasingly utilized in teledermatology, but their diagnostic value is underestimated. Recently, deep learning-based methods have received great attention in clinical skin disease image recognition and achieved promising results [3,5,11,18,20,23,25,26]. Sun *et al.* [18] released a clinical image dataset of skin diseases, namely SD-198, containing 6,584 images from 198 different categories. The results demonstrate that deep features from convolutional neural networks (CNNs) outperform handcrafted features in exploiting structural and semantic information. Gupta *et al.* [5] proposed a dual stream network that employs class activation maps to localize discriminative regions of the skin disease and exploit local features from detected regions to improve classification performance.

Although these approaches have achieved impressive results, most of them neglect the domain knowledge of dermatology and lack interpretability in diagnosis basis and results. In a typical inspection, dermatologists give an initial evaluation with the consideration of both global information, e.g. body part, and local information, e.g. the attributes of skin lesions, and further information including the patient's medical history or additional examination is required to draw a diagnostic conclusion from several possible skin diseases. Recognizing skin diseases from clinical images presents various challenges that can be summarized as follows: (1) Clinical images taken by portable electronic devices (e.g. mobile phones) often have cluttered backgrounds, posing difficulty in accurately locating lesions. (2) Skin diseases exhibit high intra-class variability in lesion appearance, but low inter-class variability, thereby making discrimination challenging. (3) The diagnostic reasoning of dermatologists is empirical and complicated, which makes it hard to simulate and model.

To tackle the above issues and leverage the domain knowledge of dermatology, we propose a novel multi-task model, namely **DermImitFormer**. The model is designed to imitate the diagnostic process of dermatologists (as shown in Fig. 1), by employing three distinct modules or strategies. Firstly, the multi-task learning strategy provides extra body parts and lesion attributes predictions, which enhances the differential diagnosis accuracy with the additional correlation from multiple predictions and improves the interpretability of diagnosis with more supporting information. Secondly, a lesion selection module is designed to imitate dermatologists' zoom-in action, effectively highlighting the local lesion features from noisy backgrounds. Thirdly, a cross-interaction module explicitly models the complicated diagnostic reasoning between body parts, lesion attributes, and diseases, increasing the feature alignments and decreasing gradient conflicts from different tasks. Last but not least, we build a new dataset containing 57,246 clinical images. The dataset includes 49 most common skin diseases, covering 80% of the consultation scenarios, 15 body parts, and 27 lesion attributes, following the International League of Dermatological Societies (ILDS) guideline [13].

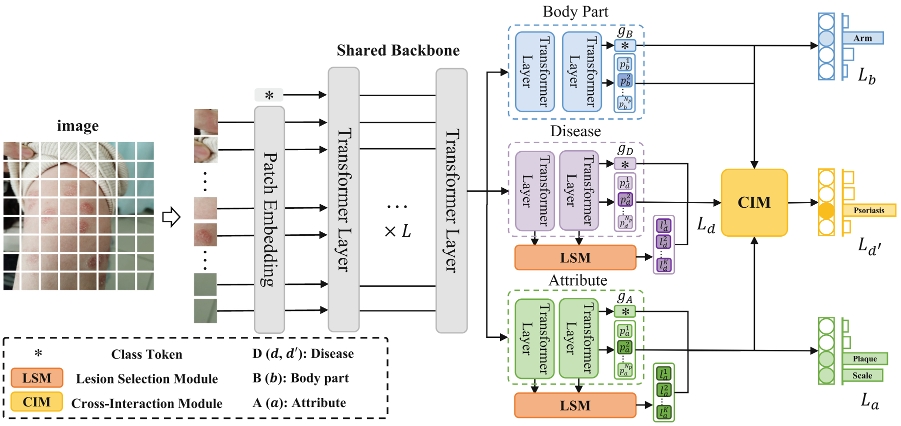


Fig. 2. The overall architecture of the multi-task imitation model (DermImitFormer) with shared backbone and task-specific heads.

The main contributions can be summarized as follows: (1) A novel multi-task model DermImitFormer is proposed to imitate dermatologists' diagnostic processes, providing outputs of diseases, body parts, and lesion attributes for improved clinical interpretability and accuracy. (2) A lesion selection module is presented to encourage the model to learn more distinctive lesion features. A cross-interaction module is designed to effectively fuse three different feature representations. (3) A large-scale clinical image dataset of skin diseases is established, containing significantly more cases than existing datasets, and closer to the real data distribution of clinical routine. More importantly, our proposed approach achieves the leading recognition performance on three different datasets.

2 Method

The architecture of the proposed multi-task model DermImitFormer is shown in Fig. 2. It takes the clinical image as input and outputs the classification results of skin diseases, body parts, and attributes in an end-to-end manner. During diagnostic processes, dermatologists consider local and global contextual features of the entire clinical image, including shape, size, distribution, texture, location, etc. To effectively capture these visual features, we use the vision transformer (ViT) [4] as the shared backbone. Three separate task-specific heads are then utilized to predict diseases, body parts, and attributes, respectively, with each head containing two independent ViT layers. In particular, in the task-specific heads of diseases and attributes, the extracted features of each layer are separated into the image features and the patch features. These two groups of features are fed into the lesion selection module (LSM), to select the most informative lesion tokens. Finally, the feature representations of diseases, body parts, and attributes are delivered to the cross-interaction module (CIM) to generate a more comprehensive representation for the final differential diagnosis.

Shared Backbone. Following the ViT model, an input image X is divided to N_p squared patches $\{x_n, n \in \{1, 2, ..., N_p\}\}$, where $N_p = (H \times W)/P^2$, P is the side length of a squared patch, H and W are the height and width of the image, respectively. Then, the patches are flattened and linearly projected into patch tokens with a learnable position embedding, denoted as $t_n, n \in \{1, 2, ..., N_p\}$. Together with an extra class token t_0, the network inputs are represented as $t_n \in \mathbb{R}^D, n \in \{0, 1, ..., N_p\}$ with a dimension of D. Finally, the tokens are fed to L consecutive transformer layers to obtain the preliminary image features.

Lesion Selection Module. As introduced above, skin diseases have high variability in lesion appearance and distribution. Thus, it requires the model to concentrate on lesion patches so as to describe the attributes and associated diseases precisely. The multi-head self-attention (MHSA) block in ViT generates global attention, weighing the informativeness of each token. Inspired by [19], we introduce a lesion selection module (LSM), which guides the transformer encoder to select the tokens that are most relevant to lesions at different levels. Specifically, for each attention head in MHSA blocks, we compute the attention matrix $\boldsymbol{A}^m = \text{Softmax}(\mathcal{QK}^T/\sqrt{D}) \in \mathbb{R}^{(N_p+1)\times(N_p+1)}$, where $m \in \{1, 2, ..., N_h\}$, N_h denoting the number of heads, $\mathcal{Q}$ and $\mathcal{K}$ the *Query* and *Key* representations of the block inputs, respectively. The first row calculates the similarities between the class token and each patch token. As the class token is utilized for classification, the higher the value, the more informative each token is. We apply softmax to the first row and the first column of $\boldsymbol{A}^m$, denoted as $a^m_{0,n}$ and $a^m_{n,0}, n \in \{1, 2, ..., N_p\}$, representing the attention scores between the class token and other tokens:

$$a_{0,n}^{m}=\frac{e^{A_{0,n}^{m}}}{\sum\limits_{i=1}^{N_p}e^{A_{0,i}^{m}}},\quad a_{n,0}^{m}=\frac{e^{A_{n,0}^{m}}}{\sum\limits_{i=1}^{N_p}e^{A_{i,0}^{m}}},\quad s_n=\frac{1}{N_h}\sum_{m=1}^{N_h}a_{0,n}^{m}\cdot a_{n,0}^{m} \tag{1}$$

The mutual attention score s_n is calculated across all attention heads. Thereafter, we select the top K tokens according to s_n for two task heads as $\boldsymbol{l}_d^k$ and $\boldsymbol{l}_a^k, k\in\{1,2,...,K\}$.

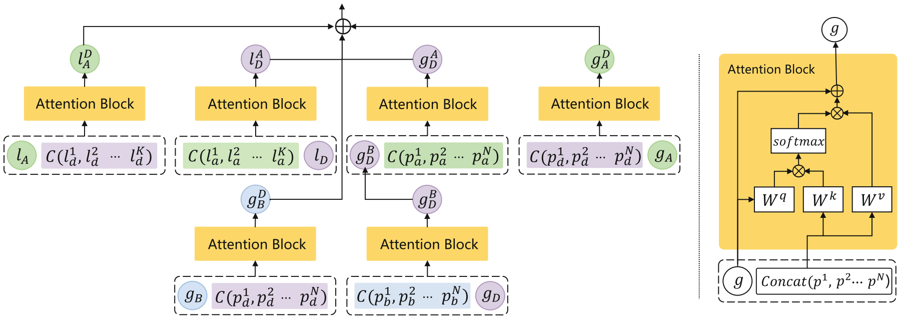

l_A^D
l_D^A
g_D^A
g_A^D
Attention Block
Attention Block
Attention Block
Attention Block
l_A
$C(l_d^1, l_d^2 \cdots l_d^K)$
$C(l_a^1, l_a^2 \cdots l_a^K)$
l_D
g_D^B
$C(p_a^1, p_a^2 \cdots p_a^N)$
$C(p_d^1, p_d^2 \cdots p_d^N)$
g_A
g_B^D
g_D^B
Attention Block
Attention Block
g_B
$C(p_d^1, p_d^2 \cdots p_d^N)$
$C(p_b^1, p_b^2 \cdots p_b^N)$
g_D
g
Attention Block
$softmax$
W^q
W^k
W^v
g
$Concat(p^1, p^2 \cdots p^N)$

Fig. 3. Schematic of cross-interaction module.

Cross-Interaction Module. A diagnostic process of skin diseases takes multiple visual information into account, which is relatively complicated and difficult to model in an analytical way. Simple fusion operations such as concatenation are insufficient to simulate the diagnostic logic. Thus, partially inspired by [21], the CIM is designed to learn complicated correlations between disease, body part, and attribute. The detailed module schematic is shown in Fig. 3. Firstly, the features of body-part and disease are integrated to enhance global representations by a cross-attention block. For example, the fusion between the class token of disease and patch tokens of body-part is:

$$z_b = LN(GAP(\boldsymbol{p}_b^1, \boldsymbol{p}_b^2, ..., \boldsymbol{p}_b^{N_p})) \tag{2}$$

$$\mathcal{Q}=LN\left(\boldsymbol{g}_D\right)\boldsymbol{W}_{BD}^{\mathcal{Q}},\quad \mathcal{K}=\boldsymbol{z}_b\boldsymbol{W}_{BD}^{\mathcal{K}},\quad \mathcal{V}=\boldsymbol{z}_b\boldsymbol{W}_{BD}^{\mathcal{V}} \tag{3}$$

$$\boldsymbol{g}_D^B=LN\left(\boldsymbol{g}_D\right)+\text{linear}(\text{softmax}(\frac{\mathcal{Q}\mathcal{K}^T}{\sqrt{F/N_h}})\mathcal{V}) \tag{4}$$

where $\boldsymbol{g}_B$, $\boldsymbol{g}_D$ are the class token, $\boldsymbol{p}_b^i$, $\boldsymbol{p}_d^i, i\in\{1,2,...,N_p\}$ the corresponding patch tokens. GAP and LN denote the global average pooling and layer normalization, respectively. $\boldsymbol{W}_{BD}^{\mathcal{Q}}, \boldsymbol{W}_{BD}^{\mathcal{K}}, \boldsymbol{W}_{BD}^{\mathcal{V}}\in\mathcal{R}^{F\times F}$ denote learnable parameters. F denotes the dimension of features. $\boldsymbol{g}_B^D$ is computed from the patch tokens of disease and the class token of body-part in the same fashion. Similarly, we can obtain the fused class tokens ($\boldsymbol{g}_A^D$ and $\boldsymbol{g}_D^A$) and the fused local class tokens ($\boldsymbol{l}_A^D$ and $\boldsymbol{l}_D^A$) between attribute and disease. Note that the disease class token $\boldsymbol{g}_D$ is replaced by $\boldsymbol{g}_D^B$ in the later computations, and local class tokens $\boldsymbol{l}_A$ and $\boldsymbol{l}_D$ in Fig. 3 are generated by GAP on selected local patch tokens from LSM. Finally, these mutually enhanced features from CIM are concatenated together to generate more accurate predictions of diseases, body parts, and attributes.

Learning and Optimization. We argue that joint training can enhance the feature representation for each task. Thus, we define a multi-task loss as follows:

Table 1. Ablation study for DermImitFormer on Derm-49 dataset. D, B, and A denote the task-specific head of diseases, body parts, and attributes, respectively.

Dimension	LSM	Fusion		F1-score (%)			Accuracy(%)
		Concat	CIM	Disease	Body part	Attribute	Disease
D				76.2	-	-	80.4
D	✓			77.8	-	-	82.0
D + B	✓	✓		78.1	85.0	-	82.4
D + A	✓	✓		78.4	-	68.7	82.6
D + B + A	✓	✓		79.1	85.1	69.0	82.9
D + B + A	✓		✓	**79.5**	**85.9**	**70.4**	**83.3**

$$\mathcal{L}_x = -\frac{1}{N_s}\sum_{i=1}^{N_s}\sum_{j=1}^{n_x} y_{ij}\log(p_{ij}), \quad x \in \{d, d'\} \tag{5}$$

$$\mathcal{L}_h = -\frac{1}{N_s}\sum_{i=1}^{N_s}\sum_{j=1}^{n_h} y_{ij}\log(p_{ij}) + (1-y_{ij})\log(1-p_{ij}), \quad h \in \{a, b\} \tag{6}$$

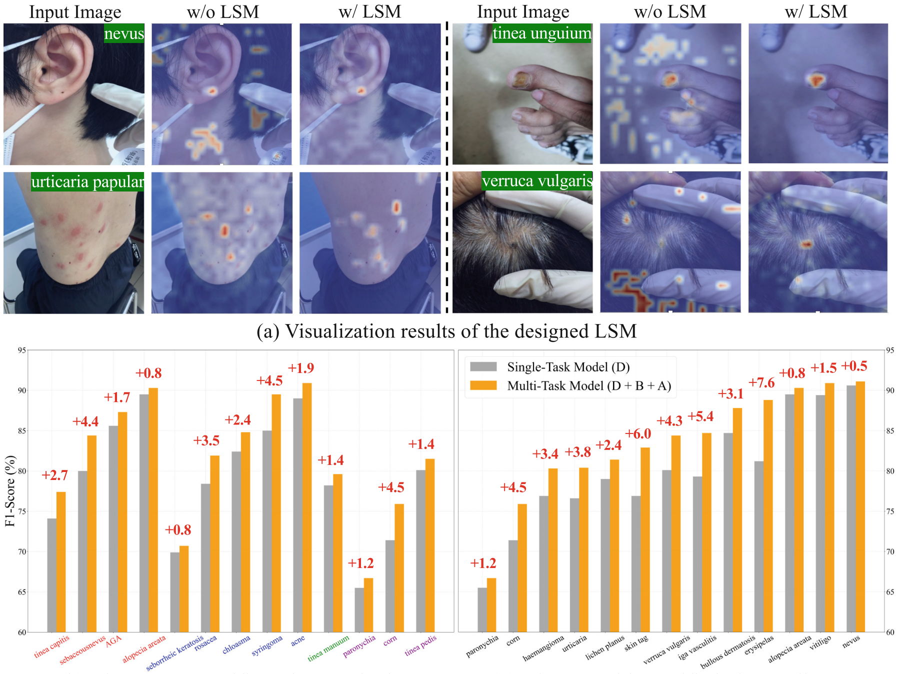
$$\mathcal{L} = \mathcal{L}_d + \mathcal{L}_{d'} + \mathcal{L}_b + \mathcal{L}_a \tag{7}$$

where N_s denotes the number of samples, n_x, n_h the number of classes for each task, and p_{ij}, y_{ij} the prediction and label, respectively. Notably, body parts and attributes are defined as multi-label classification tasks, optimized with the binary cross-entropy loss, as shown in Eq. 6. The correspondence of x and h is shown in Fig. 2.

3 Experiment

Datasets. The proposed DermImitFormer is evaluated on three different clinical skin image datasets including an in-house dataset and two public benchmarks. (1) **Derm-49:** We establish a large-scale clinical image dataset of skin diseases, collected from three cooperative hospitals and a teledermatology platform. The 57,246 images in the dataset were annotated with the diagnostic ground truth of skin disease, body parts, and lesions attributes from the patient records. We clean up the ground truth into 49 skin diseases, 15 body parts, and 27 lesion attributes following the ILDS guidelines [13]. (2) **SD-198** [18]**:** It is one of the largest publicly available datasets in this field containing 198 skin diseases and 6,584 clinical images collected through digital cameras or mobile phones. (3) **PAD-UFES-20** [15]**:** The dataset contains 2,298 samples of 6 skin diseases. Each sample contains a clinical image and a set of metadata with labels such as diseases and body parts.

Implementation Details. The DermImitFormer is initialized with the pre-trained ViT-B/16 backbone and optimized with SGD method (initial learning rate 0.003, momentum 0.95, and weight decay 10^{-5}) for 100 epochs on 4 NVIDIA Tesla V100 GPUs with a batch size of 96. We define the input size i.e. $H = W = 384$ that produces a total of 576 spatial tokens i.e. $N_p = 576$ for a ViT-B backbone. $K = 24$ in the LSM module. For data augmentation, we employed the Cutmix [24] with a probability of 0.5 and Beta(0.3, 0.3) during optimization. We adopt precision, recall, F1-score, and accuracy as the evaluation metrics.



(a) Visualization results of the designed LSM

(b) Diseases on specific pathogenesis sites

(c) Diseases with specific lesion attributes

Fig. 4. Comparative results on Derm-49 dataset. Red, blue, green, and purple fonts denote diseases on heads, faces, hands, and feet (best viewed in color).

Ablation Study. The experiment is conducted based on ViT-B/16 and the results are reported in Table 1. (1) **LSM:** Quantitative results demonstrate that the designed LSM yields 1.6% improvement in accuracy. Qualitative results are shown in Fig. 4(a), which depicts the attention maps obtained from the last transformer layer. Without LSM, vision transformers would struggle of localizing lesions and produce noisy attention maps. With LSM, the attention maps are more discriminative and lesions are localized precisely, regardless of variations in terms of scale and distribution. (2) **Multi-task learning:** The models are trained with a shared backbone and different combinations of task-specific heads. The results show that multi-task learning (D+B+A) increases the F1-score from 77.8 to 79.1. (3) **CIM:** Quantitative results show that the presented CIM can further improve the F1-score of diseases to 79.5. Notably, the *p*-value of 1.03e-05 (<0.01) is calculated by comparing the results of 5-fold cross-validation with the baseline, illustrating the significance of our model. In particular, the representation with fused features of body parts and attributes can improve the recognition performance of diseases. As shown in Fig. 4(b) and (c), statistics show that the classification performance of these diseases is improved by the multi-task learning strategy and CIM. For instance, rosacea and tinea versicolor share the same attributes of macule and papular, but rosacea typically affects the face. By fusing the representation of body parts, the F1-score of rosacea is increased by 4.5%. Similarly, our model improves the recognition accuracy of diseases with distinctive lesion attributes such as skin tags, urticaria, etc. Meanwhile, the extra information about body parts and attributes improves the interpretability of diagnoses.

Table 2. Comparison to state-of-the-art methods on Derm-49 dataset (top) and two public benchmarks: SD-198 dataset (mid), PAD-UFES-20 dataset (bottom).

Datasets	Methods	F1-score (%)	Precision (%)	Recall (%)	Accuracy (%)
Derm-49	DX [6]	72.6 ± 2.3	73.7 ± 0.6	72.2 ± 3.1	73.4 ± 0.7
	ViT-Base [4]	75.9 ± 0.8	80.6 ± 0.7	72.9 ± 0.9	80.4 ± 0.4
	Swin-Base [12]	76.6 ± 0.6	83.5 ± 1.1	71.0 ± 0.9	80.6 ± 0.5
	DermImitFomer	**78.8 ± 0.5**	**83.5 ± 0.6**	**74.6 ± 1.1**	**82.6 ± 0.5**
SD-198	SPBL [23]	66.2 ± 1.6	71.4 ± 1.7	65.7 ± 1.6	67.8 ± 1.8
	Aux-D [22]	68.0 ± 1.0	67.9 ± 1.0	69.2 ± 0.9	-
	Dual Stream [5]	70.9 ± 1.2	73.1 ± 1.4	69.2 ± 1.1	71.4 ± 1.1
	TPC [9]	63.2 ± 1.6	65.6 ± 1.7	64.7 ± 1.6	-
	IASN [3]	68.6 ± 0.7	71.9 ± 0.8	70.0 ± 0.9	70.7 ± 0.8
	PCCT [2]	65.2 ± 1.6	68.4 ± 1.4	66.0 ± 1.5	-
	DermImitFomer-ST	**73.6 ± 2.6**	**76.1 ± 2.6**	**75.1 ± 2.2**	**74.5 ± 2.6**
PAD-UFES-20	PAD [15]	71.0 ± 2.9	73.4 ± 2.9	70.8 ± 2.8	70.7 ± 2.8
	T-Enc [14]	-	-	-	61.6 ± 5.1
	ResNet-50 [15]	67.8 ± 3.7	72.0 ± 4.1	67.0 ± 4.1	67.1 ± 4.1
	ViT-Base [4]	69.9 ± 1.4	69.4 ± 1.5	70.4 ± 2.2	70.6 ± 1.8
	Swin-Base [12]	72.1 ± 2.5	72.0 ± 2.9	72.7 ± 2.6	72.7 ± 2.5
	DermImitFomer-ST	73.6 ± 2.8	72.8 ± 3.2	74.4 ± 2.4	74.4 ± 2.4
	DermImitFomer	**74.5 ± 2.5**	**73.9 ± 2.9**	**75.0 ± 2.1**	**75.0 ± 2.1**

Results. To evaluate the effectiveness of our proposed DermImitFormer, we conduct a comparison with various state-of-the-art methods on three different datasets. The results are reported in Table 2. (1) **Derm-49:** Compared with other state-of-the-art approaches, our proposed DermImitFormer achieves the leading classification performance in our established dataset with the 5-fold cross-validation splits. (2) **SD-198:** Since the dataset does not contain labels of lesion attributes and body parts, the proposed DermImitFormer in Single-Task mode (w/o CIM) is implemented in the experiment. The result is based on the provided 5-fold cross-validation splits. Quantitative results in Table 2(mid) demonstrate that our proposed DermImitFormer-ST achieves state-of-the-art classification performance. In contrast to other approaches, our model can precisely localize more discriminative lesion regions and thus has superior classification accuracy. (3) **PAD-UFES-20:** The dataset contains labels of diseases and body parts. Thus, the proposed DermImitFormer with different modes is evaluated in the experiment by the 5-fold cross-validation splits. Quantitative results in Table 2 (bottom) demonstrate that our proposed model outperforms the CNN-based [14,15], and transformer-based methods [4,12], achieving the state-of-the-art classification performance. In particular, the performance of DermImitFormer is better than that of DermImitFormer-ST in Single-Task mode (w/o CIM), which further indicates the effectiveness of the multi-task learning strategy and CIM.

4 Conclusion

In this work, DermImitFormer, a multi-task model, has been proposed to better utilize dermatologists' domain knowledge by mimicking their subjective diagnostic procedures. Extensive experiments demonstrate that our approach achieves state-of-the-art recognition performance in two public benchmarks and a large-scale in-house dataset, which highlights the potential of our approach to be employed in real clinical environments and showcases the value of leveraging domain knowledge in the development of machine learning models.

Acknowledgement. This work was supported by National Key R&D Program of China (2020YFC2008703) and the Project of Intelligent Management Software for Multimodal Medical Big Data for New Generation Information Technology, the Ministry of Industry and Information Technology of the People's Republic of China (TC210804V).

References

1. Binder, M., et al.: Epiluminescence microscopy: a useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists. Arch. Dermatol. **131**(3), 286–291 (1995)
2. Chen, K., Lei, W., Zhang, R., Zhao, S., Zheng, W.S., Wang, R.: PCCT: progressive class-center triplet loss for imbalanced medical image classification. arXiv preprint arXiv:2207.04793 (2022)

3. Chen, X., Li, D., Zhang, Y., Jian, M.: Interactive attention sampling network for clinical skin disease image classification. In: Ma, H., et al. (eds.) PRCV 2021. LNCS, vol. 13021, pp. 398–410. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-88010-1_33
4. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Gupta, K., Krishnan, M., Narayanan, A., Narayan, N.S., et al.: Dual stream network with selective optimization for skin disease recognition in consumer grade images. In: Proceedings of the International Conference on Pattern Recognition (ICPR), pp. 5262–5269. IEEE (2021)
6. Jalaboi, R., Faye, F., Orbes-Arteaga, M., Jørgensen, D., Winther, O., Galimzianova, A.: Dermx: an end-to-end framework for explainable automated dermatological diagnosis. Med. Image Anal. **83**, 102647 (2023)
7. Kittler, H., Pehamberger, H., Wolff, K., Binder, M.: Diagnostic accuracy of dermoscopy. Lancet Oncol. **3**(3), 159–165 (2002)
8. Kshirsagar, P.R., Manoharan, H., Shitharth, S., Alshareef, A.M., Albishry, N., Balachandran, P.K.: Deep learning approaches for prognosis of automated skin disease. Life **12**(3), 426 (2022)
9. Lei, W., Zhang, R., Yang, Y., Wang, R., Zheng, W.S.: Class-center involved triplet loss for skin disease classification on imbalanced data. In: Proceedings of the International Symposium on Biomedical Imaging (ISBI), pp. 1–5. IEEE (2020)
10. Li, L.F., Wang, X., Hu, W.J., Xiong, N.N., Du, Y.X., Li, B.S.: Deep learning in skin disease image recognition: a review. IEEE Access **8**, 208264–208280 (2020)
11. Liu, Y., et al.: A deep learning system for differential diagnosis of skin diseases. Nat. Med. **26**(6), 900–908 (2020)
12. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022 (2021)
13. Nast, A., Griffiths, C.E., Hay, R., Sterry, W., Bolognia, J.L.: The 2016 international league of dermatological societies' revised glossary for the description of cutaneous lesions. Br. J. Dermatol. **174**(6), 1351–1358 (2016)
14. Ou, C., et al.: A deep learning based multimodal fusion model for skin lesion diagnosis using smartphone collected clinical images and metadata. Front. Surg. **9**, 1029991 (2022)
15. Pacheco, A.G., Krohling, R.A.: The impact of patient clinical information on automated skin cancer detection. Comput. Biol. Med. **116**, 103545 (2020)
16. Rogers, H.W., Weinstock, M.A., Feldman, S.R., Coldiron, B.M.: Incidence estimate of nonmelanoma skin cancer (keratinocyte carcinomas) in the us population, 2012. JAMA Dermatol. **151**(10), 1081–1086 (2015)
17. Siegel, R.L., Miller, K.D., Fuchs, H.E., Jemal, A.: Cancer statistics, 2022. CA Cancer J. Clin. **72**(1), 7–33 (2022)
18. Sun, X., Yang, J., Sun, M., Wang, K.: A benchmark for automatic visual classification of clinical skin disease images. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 206–222. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_13
19. Wang, J., Yu, X., Gao, Y.: Feature fusion vision transformer for fine-grained visual categorization. arXiv preprint arXiv:2107.02341 (2021)
20. Wu, J., et al.: Learning differential diagnosis of skin conditions with co-occurrence supervision using graph convolutional networks. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12262, pp. 335–344. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59713-9_33

21. Xu, J., et al.: Remixformer: a transformer model for precision skin tumor differential diagnosis via multi-modal imaging and non-imaging data. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13433, pp. 624–633. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16437-8_60
22. Xu, Z., Zhuang, J., Zhang, R., Wang, R., Guo, X., Zheng, W.S.: Auxiliary decoder and classifier for imbalanced skin disease diagnosis. J. Phys. Conf. Ser. **1631**(1), 012046 (2020)
23. Yang, J., et al.: Self-paced balance learning for clinical skin disease recognition. IEEE Trans. Neural Netw. Learn. Syst. **31**(8), 2832–2846 (2019)
24. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6023–6032 (2019)
25. Zhang, J., Xie, Y., Wu, Q., Xia, Y.: Medical image classification using synergic deep learning. Med. Image Anal. **54**, 10–19 (2019)
26. Zhang, J., Xie, Y., Xia, Y., Shen, C.: Attention residual learning for skin lesion classification. IEEE Trans. Med. Imaging **38**(9), 2092–2103 (2019)