RemixFormer: A Transformer Model for Precision Skin Tumor Differential Diagnosis via Multi-modal Imaging and Non-imaging Data

Jing Xu[1], Yuan Gao[1], Wei Liu[1], Kai Huang[2], Shuang Zhao[2], Le Lu[1]
Xiaosong Wang[1], Xian-Sheng Hua[1], Yu Wang[1(✉)], and Xiang Chen[2]

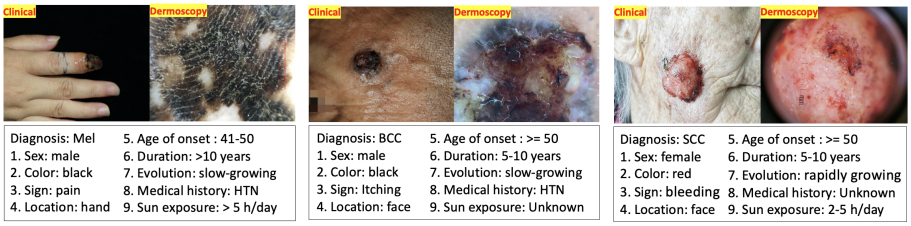
[1] DAMO Academy, Alibaba Group, Hangzhou, China
tonggou.wangyu@alibaba-inc.com
[2] Department of Dermatology, Xiangya Hospital Central South University, Changsha, China

Abstract. Skin tumor is one of the most common diseases worldwide and the survival rate could be drastically increased if the cancerous lesions were identified early. Intrinsic visual ambiguities displayed by skin tumors in multi-modal imaging data impose huge amounts of challenges to diagnose them precisely, especially at the early stage. To achieve high diagnosis accuracy or precision, all possibly available clinical data (imaging and/or non-imaging) from multiple sources are used, and even the missing-modality problem needs to be tackled when some modality may become unavailable. To this end, we first devise a new disease-wise pairing of all accessible patient data if they fall into the same disease category as a remix operation of data samples. A novel cross-modality-fusion module is also proposed and integrated with our transformer-based multi-modality deep classification framework that can effectively perform multi-source data fusion (i.e., clinical images, dermoscopic images and accompanied with clinical patient-wise metadata) for skin tumors. Extensive quantitative experiments are conducted. We achieve an absolute 6.5% increase in averaged F1 and 2.8% in accuracy for the classification of five common skin tumors by comparing to the prior leading method on Derm7pt dataset of 1011 cases. More importantly, our method obtains an overall 88.5% classification accuracy using a large-scale in-house dataset of 5601 patients and in ten skin tumor classes (pigmented and non-pigmented). This experiment further validates the robustness and implies the potential clinical usability of our method, in a more realistic and pragmatic clinic setting.

Keywords: Skin Tumor · Multi-modality Fusion · Remix Sampling

1 Introduction

Accurate early skin lesion diagnosis is crucial to prevent skin cancers and can significantly increase the 5-year survival rate of malignant tumors [10] such as Melanoma (Mel). However, it remains a challenging task even for well-trained



(a) Mel (b) BCC (c) SCC

Fig. 1: Samples of X-SkinTumor-10 dataset

professionals, and experienced dermatologists find clues in multi-modal data from different types of clinical sources. For example, precision diagnosis of skin tumors often involves examining lesions, analyzing details in dermatoscopic images, and referencing metadata, e.g., medical history. Fig. 1 shows three skin tumor cases with different conditions, where the clinical and dermoscopic images are listed and accompanied with patients' meta information. Specifically, in Fig. 1(a), patient's non-imaging data (the age, slow progression over the years, and long duration of the sun exposure) largely support the diagnosis of malignant melanoma in addition to the observation of black plaque visually.

In recent years, computer-aided diagnosis has shown some impressive performance in supporting dermatologists' diagnoses [4–7, 11–15]. Esteva et al. [4] demonstrated that the convolutional neural networks (CNN) are capable of classifying malignant and benign skin tumors with a level of competence comparable to dermatologists. Tschandl et al. [13, 14] also reported that learning-based classifiers not only can outperform human experts in the diagnosis of pigmented skin lesions but can further improve the diagnostic accuracy when high quality computer-aided clinical decision-making is available to dermatologists. Such a way performs better over those by either algorithms or physicians alone. When it comes to employing multi-modal data, Ge et al. [5] proposed a deep convolutional neural network architecture to capture discriminative features from both clinical and dermoscopic images and showed that the multi-modality method significantly outperforms single-modality methods. Furthermore, Haenssle et al. [6, 7] studied CNN's diagnostic performance with a large group of dermatologists, in which most dermatologists were outperformed by the CNN models when only dermoscopic images are provided. When given multi-modal information, most dermatologists only performed equivalently as the CNN-based models. Thus, one can reasonably assume that better performance can be achieved provided that the neural networks are also trained with a range of multi-modal data as dermatologists did. Most recently, Tang et al. [12] constructed a two-stage approach named FusionM4Net. It concatenates features of clinical and dermoscopic images at the first stage, and then incorporates the patient's metadata with the prediction from the first stage via SVM-based clustering. The final diagnosis is formed by the fusion of the predictions from two stages.

To achieve the performance on a par with human experts, a computer-aided skin tumor diagnosis system should have the similar capability of processing

multi-modal data, as a dermatologist does in a realistic clinic environment. However, training neural network on multi-modal data also add extra burden to the already tidy data collection and annotation process. Although we can adopt and use the data as we have, we face several technical challenges while training the multi-modal neural networks. First, it is more than common in multi-modal data to have some modalities inaccessible, whenever they are not acquired or indeed missing. It is often unrealistic to ensure the completeness of every modality for each sample/patient. Furthermore, another difficulty is how to effectively merge the multiple source of information for the multi-modality models, especially together with the first challenge in the training phase. In contrast to previous methods [8, 1, 12], where paired data are required for the training, a novel data sampling strategy via disease-wise pairing (DWP as a remix of data samples on the disease-class level) is presented. Integrated with a scalable cross-modality-fusion module, our proposed multi-modal classification framework can better handle the incomplete data in the model training and achieve higher classification accuracy in both the publicly released dataset (covering mainly pigmented tumors) and a large-scale private skin tumor dataset with 10 categories, containing additional non-pigmented skin tumors such as basal cell carcinoma (BCC) and squamous cell carcinoma (SCC) which have higher incidence rates in clinics.

Our contributions are three-fold: (a) We propose a new transformer-based multi-modality classification framework for skin tumors, which includes the novel DWP sampling strategy for tackling the missing modality issue in training data. This disease-wise pairing augmentation process makes our framework more flexible for model training, and more importantly, provides better training generalizability for the trained model; (b) We improve the multi-modal data fusion with a more efficient cross-modality fusion module than the conventional concatenation. A leading recognition accuracy of 81.3% is achieved on Derm7pt [8] dataset; (c) We compose a large-scale multi-modality dataset with significantly more cases than the existing databases and importantly, it covers both pigmented and non-pigmented skin tumors (closer to the real data distributions of daily clinical routines). Our multi-modal framework achieves 88.5% accuracy and 98.4% AUC on this more comprehensive and realistic database.

2 Method

Our proposed framework is outlined in the Fig. 2. During training, the multi-modal samples are first grouped in a disease-wise fashion, i.e., multi-modal data can be randomly selected and paired as long as they belong to the same disease category. It can be seen as a remix of the multi-modal data on the disease level since a multi-modal tuple here will have data from different studies/patients. There are numerous possible permutations of modalities among all the data in the same disease class. It not only significantly increases the amount of training data (as a form of effective data augmentation) but also largely enhances the robustness and generalizability of the trained model.

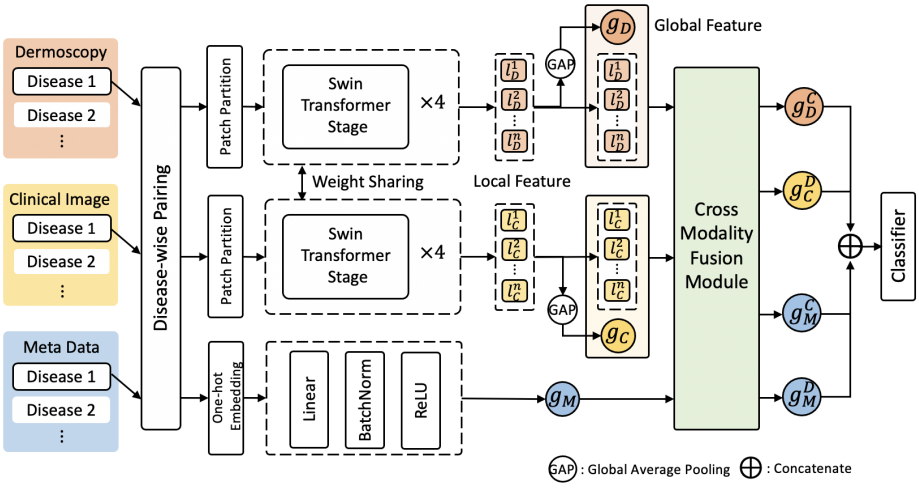


Fig. 2: The pipeline of our multi-modal cross-fusion transformer

A Swin Transformer [9] is adopted as the unified backbone to encode the clinical and dermoscopic images. Metadata is separately processed using the one-hot representation and followed by a linear embedding. The image features and metadata feature are fed into the cross-modality fusion (CMF) module to form a global representation for the final disease classification. The entire framework can be trained in an end-to-end manner using a standard cross entropy loss.

2.1 Unified Transformer Backbone

During diagnosis, experienced dermatologists usually consider various visual features including shape, size, color, texture, location and distribution, etc., of which some are localized features and some need spatial global context. Although multi-scale convolutional neural networks can be used to model these complicated local-global interactions, Transformer [3] based methods (ViTs) are more suitable choices by the virtue of their non-local modeling capabilities. Besides, Transformer is a natural choice for multi-modal feature encoding. As an improvement to ViT, Swin Transformer [9] only computes self-attention locally within non-overlapping windows and thus has less complexity. From previous work [8, 1, 12], separate backbones are employed for clinical and dermoscopic images. We empirically find that using the shared backbone for both clinical and dermoscopic images will not compromise the performance, and actually this simplicity design of choice makes the training and inference more efficient. Taking all these considerations into account, a unified Swin Transformer backbone is adopted into our multi-modal framework.

2.2 Disease-wise Pairing

Missing modality has been common when dealing with multi-modal data. Requiring the completeness of every modality in the clinical dataset adds extra

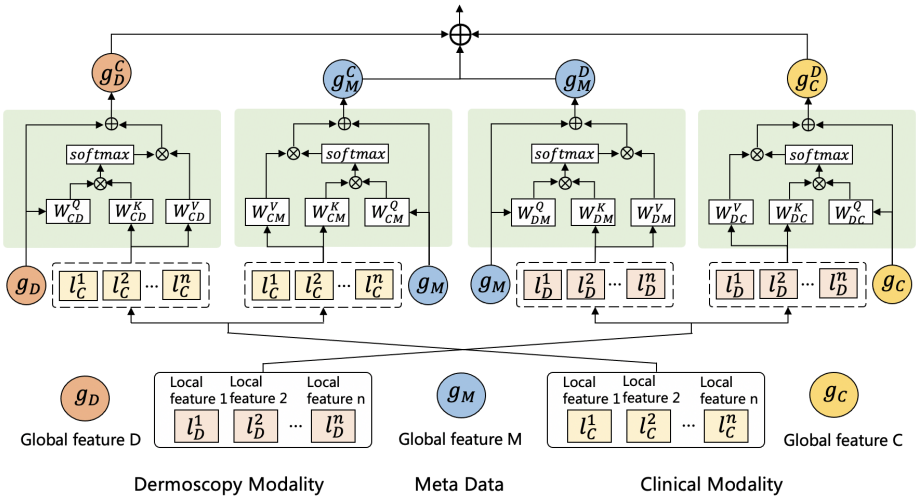


Fig. 3: The Cross-Modality-Fusion module.

burden to the already hash process of data collection and cleanse. To learn a good multi-modality representation at the class level, we integrate a disease-wise pairing scheme into our data augmentation pipeline during training to complete some of the modality missing. Unlike using instance-wise pairs for training, the DWP encourages the network to input/fuse the different modality data channels from different patients (as long as they belong under the same disease category) into the feature space. Our framework is convenient to add new training samples for any modality, which greatly reduces the cost of data collection and maximizes the data utilization. Furthermore, it serves as an additional regularization to prevent the model from learning some spurious correlation within a certain pair of modalities, and thus provides better generalizability (i.e., semantic data modality fusion at the disease level) and extra performance boosts.

Specifically, we define a sample data as $s_i = \{c_i, d_i, m_i\}$, where c_i, d_i, and m_i represent a clinical image, a dermoscopic image, and a set of metadata respectively. $S = \{s_i \mid i \in \{1, ..., I\}\}$ is used to indicate the set of sampled training data with the total amount of I. Let C, D, M be the three modalities of clinical image, dermoscopy and metadata, respectively; C^k, D^k, M^k be the grouped sets of data with the corresponding disease class k. When DWP is turned on (based on $p > T_p, p \in [0, 1]$), c_i is randomly sampled from $\{c_j^k \mid \forall\ c_j^k \in C^k, k \in \{1, ..., K\}\}$, d_i from $\{d_j^k \mid \forall\ d_j^k \in D^k, k \in \{1, ..., K\}\}$, and m_i from $\{m_j^k \mid \forall\ m_j^k \in M^k, k \in \{1, ..., K\}\}$, where K is the number of diseases to be considered. Otherwise, $\{c_i, d_i, m_i\}$ will naturally be from the same study and patient if available (zero padding is used for missing modalities during training and testing). T_p is a cut-off threshold to control the probability of applying DWP to the input samples, where p is a random number generated by a uniform distribution on the interval [0, 1]. For each input sample, we apply DWP when $p > T_p$ (T_p is empirically set to 0.6).

2.3 Cross-Modality Fusion

Partially inspired by [2], the CMF module is designed to fuse the global features of each modality with the local features from another modality in a cyclic manner across all modalities. Specifically, the global features from each modality will exchange information with the local features from other modalities through a multi-head attention module. Since the global feature has already gathered the information from local features in its own modality, the multi-head attention will fuse the information from the local features of other modalities.

The detailed diagram for the fusion of three engaged modalities is shown in Fig. 3. $\boldsymbol{l}_C$ or $\boldsymbol{l}_D$ is the output feature map of the last stage in Swin Transformer. $\boldsymbol{g}_C$ or $\boldsymbol{g}_D$ are generated by applying a global average pooling (GAP) layer, and they are the class tokens for ViT backbones. $\boldsymbol{g}_M$ is the output by a linear layer after one-hot embedding with metadata. After layer normalization (LN), a cross-attention block is utilized to fuse features by taking the local features as $\mathcal{K}$ and $\mathcal{V}$ and global features as $\mathcal{Q}$:

$$\boldsymbol{g}_{X'} = \begin{cases} \text{GAP}(\boldsymbol{l}^1_{X'}, \boldsymbol{l}^2_{X'}, ..., \boldsymbol{l}^n_{X'}), & X' \in \{C, D\} \\ \boldsymbol{g}_M, & X' = M \end{cases} \tag{1}$$

$$\boldsymbol{f}^X_{X'} = \text{LN}\left(concat(\boldsymbol{g}_{X'}, \boldsymbol{l}^1_X, \boldsymbol{l}^2_X, ..., \boldsymbol{l}^n_X)\right), \quad \tilde{\boldsymbol{g}}_{X'} = \boldsymbol{f}^X_{X'}[0], \quad \boldsymbol{z}_X = \boldsymbol{f}^X_{X'}[1:] \tag{2}$$

$$\mathcal{Q} = \tilde{\boldsymbol{g}}_{X'} \boldsymbol{W}^{\mathcal{Q}}_{XX'}, \quad \mathcal{K} = \boldsymbol{z}_X \boldsymbol{W}^{\mathcal{K}}_{XX'}, \quad \mathcal{V} = \boldsymbol{z}_X \boldsymbol{W}^{\mathcal{V}}_{XX'} \tag{3}$$

$$\boldsymbol{M}_{att} = \text{softmax}(\frac{\mathcal{Q}\mathcal{K}^T}{\sqrt{F/h}}), \quad \boldsymbol{M}_{cross} = \boldsymbol{M}_{att}\mathcal{V} \tag{4}$$

$$\boldsymbol{g}^X_{X'} = \tilde{\boldsymbol{g}}_{X'} + \text{linear}(\boldsymbol{M}_{cross}), \quad X \in \{C, D\}, \quad X' \in \{C, D, M\} \setminus X \tag{5}$$

where X and X' are defined as two different modalities, $\boldsymbol{l}$ is the local feature and $\boldsymbol{g}$ is the global feature. $\boldsymbol{W}^{\mathcal{Q}}_{XX'}, \boldsymbol{W}^{\mathcal{K}}_{XX'}, \boldsymbol{W}^{\mathcal{V}}_{XX'} \in \mathcal{R}^{F \times F}$ are learnable parameters, F is the dimension of features and h is the number of heads.

3 Experiment and Results

Data We employ two datasets in this study: one public and one private dataset. The public dataset Derm7pt contains 413 training cases, 203 validation cases and 395 testing cases. Each case comprises a dermoscopic image and a clinical image, the diagnostic label is divided into 5 types: Mel, Nevus (Nev), Seborrheic Keratosis (SK), BCC, and Miscellaneous (Misc). Please refer to [8] for details, on which we will compare our method with the previous leading Fusion4MNet [12]. Our private dataset named X-SkinTumor-10 was collected from Xiangya Hospital from 2016 to 2021, and annotated by the dermatologists with at least five years experience. The dataset contains 14,941 images and 5,601 patients, and only 2,198 patients have the three-modality paired data. The percentages of missing data are 0.4%, 33.6% and 37.7% for clinical images, dermoscopic images and metadata, respectively. The patient's metadata with 9 attributes is shown in

Method	F1						Acc
	Nev	BCC	Mel	Misc	SK	Avg.	
HcCNN[1]	—	—	0.605	—	—	—	0.699
Inception-comb[8]	0.838	0.588	0.632	0.559	**0.552**	0.634	0.742
ViT-S/16 [3]	0.855	0.414	0.723	0.706	0.429	0.625	0.777
FusionM4net[12]	0.878	0.452	0.727	0.646	0.181	0.577	0.785
RemixFormer	**0.883**	**0.595**	**0.755**	**0.743**	0.519	**0.699**	**0.813**

Table 1: Comparison of our proposed model with other methods on Derm7pt.

Fig. 1. The dataset has ten types of skin tumors: SCC (8.5%), Mel (6.1%), BCC (13.4%), AK (2%), keloid (Kel, 3.4%), dermatofibroma (DF, 3.2%), sebaceous nevus (SN, 3.1%), SK (22.9%), Nev (35.7%), haemangioma (Hem, 1.6%).

Implementations On X-SkinTumor-10, we perform 5-fold cross-validation to evaluate our method, where the entire dataset is randomly divided into 5 folds on the patient level with a 3:1:1 ratio as training, validation, and testing set. Our backbone is a regular Swin-B/384, the data augmentation includes flip, rotation, random affine transformation. We adopt SGD optimizer with cosine learning rate schedule, and the initial learning rate is 1e-4. Models are trained for 200 epochs on 4 NVIDIA Tesla V100 GPUs with batch size 64. On Derm7pt, we use exactly the same data augmentation as [12], and apply Swin-T/224 model with fewer parameters to verify the effectiveness of our method. In the CMF module, the dimensions of the local features and global features are 768 in Swin-T, 1024 in Swin-B, and the number of heads is 8. We initialize the model parameters with ImageNet pretrained weights to speed up the convergence. We utilize area under the curve, macro-averaged F1-score, sensitivity, precision, specificity, and overall accuracy as the evaluation metrics, corresponding to the abbreviations AUC, F1, Sen, Pre, Spe, and Acc respectively.

Results As listed in Table 1, we compare the proposed RemixFormer with other methods on the public dataset of Derm7pt [8]. For completeness, we use ViT-S/16 and fuse the three features (only class tokens of C and D, metadata feature) by concatenation. Our method with Swin-T backbone outperforms FusionM4Net by 12.2% and 2.8% in the average F1 and overall Acc, respectively. The relatively poor performance of SK is mainly due to the long-tail problem, which may be addressed in future work. It worth mentioning that our model has noticeably less parameters than FusionM4Net (32.3M vs. 54.4M).

Using 5-fold cross-validation, RemixFormer with Swin-B backbone achieves an overall 88.5%±0.8% classification accuracy on the more comprehensive X-SkinTumor-10. The average AUC, F1, Sen, Spe and Pre of the 10 conditions are 98.4%±0.3%, 81.2%±1.2%, 80.4%±2.2%, 98.6%±0.1% and 82.8%±1.1%, respectively, and the corresponding metrics for each condition are shown in Fig. 4. The huge performance difference between Nev and Hem is largely due to the

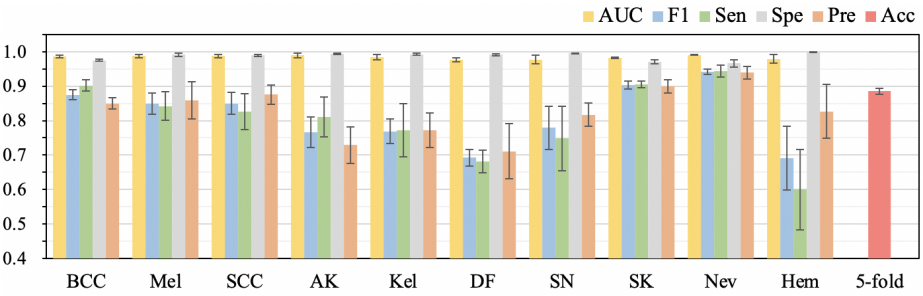


Fig. 4: 5-fold cross validation result on X-SkinTumor-10. The far right is the average accuracy of 5 folds, and error bars indicate the standard deviation.

Modality	Fusion		DWP	AUC	F1	Sen	Pre	Spe	Acc
	concat	CMF							
Clinical				0.952	0.599	0.592	0.619	0.968	0.766
Dermoscopy				0.953	0.606	0.601	0.629	0.769	0.778
C + D	✓			0.962	0.661	0.717	0.630	0.974	0.788
C + D		✓		0.966	0.681	0.729	0.652	0.975	0.805
C + D + M	✓			0.970	0.702	0.744	0.677	0.979	0.835
C + D + M		✓		0.976	0.749	0.765	0.740	0.983	0.864
C + D + M		✓	✓	**0.984**	**0.784**	**0.795**	**0.778**	**0.985**	**0.885**

Table 2: Ablation Study for multi-modality models.

imbalanced data distribution. For multi-instance cases, we use majority voting to select the prediction during inference.

To justify our design choices, we set aside 1034 three-modality cases from X-SkinTumor-10 as the test set to perform ablation study. Each case has one clinical and one dermoscopic image with metadata. We first conduct the modality-wise ablation experiments on this fixed test set. As listed in Table 2, model using only dermoscopic images performs better than clinical images. Combining two modalities (C+D) brings better performance, and simple features concatenation can improve F1 and Acc by 5.5% and 1% respectively. When the concatenation is replaced by CMF, the F1 and Acc are further improved by 2% and 1.7%. Increasing the modality to three (C+D+M) also brings significant performance gain. The model with CMF again outperforms the one using concatenation, with 4.7% and 2.9% improvements in F1 and Acc. We further validated the effectiveness of CMF with a t-test by running ten times, and the p-values of the F1 and Acc are 1.8e-4 and 7.2e-6 ($p < 0.01$). Additionally, DWP well addresses the missing modality problem and provides extra performance gain, achieving an overall accuracy of 88.5%.

4 Conclusion

We have presented an effective multi-modal transformer model, in which multi-modal data are properly fused by a novel cross-modality fusion module. To handle the missing modality problem, we implement a new disease-wise sampling strategy, which augments to form class-wise multi-modality image pairs (within the same class) and facilitates sufficient training. Swin transformer as an efficient image feature extractor is shared by both the dermoscopic and clinical image network streams. Through quantitative experiments, we achieve a new record on the public dataset of Derm7pt, surpassing the previous best method by 2.8% in accuracy. Our method is also validated on a large scale in-house dataset X-SkinTumor-10 where our reported quantitative performance of an overall 88.5% classification accuracy on recognizing ten classes of pigmented and non-pigmented skin tumors demonstrates excellent clinical potential.

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. Bi, L., Feng, D.D., Fulham, M., Kim, J.: Multi-label classification of multi-modality skin lesion via hyper-connected convolutional neural network. Pattern Recognition **107**, 107502 (2020)
2. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 357–366 (2021)
3. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
4. Esteva, A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(115–118) (2017)
5. Ge, Z., Demyanov, S., Chakravorty, R., Bowling, A., Garnavi, R.: Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images. In: Medical Image Computing and Computer Assisted Intervention (MICCAI 2017). pp. 250–258 (2017)
6. Haenssle, H.A., et al.: Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Annals of Oncology **29**, 1836–1842 (2018)
7. Haenssle, H.A., et al.: Man against machine reloaded: performance of a market-approved convolutional neural network in classifying a broad spectrum of skin lesions in comparison with 96 dermatologists working under less artificial conditions. Annals of Oncology **31**, 137–143 (2020)
8. Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G.: Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE Journal of Biomedical and Health Informatics **23**(2), 538–546 (2019)

9. 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)
10. Perera, E., Gnaneswaran, N., Jennens, R., Sinclair, R.: Malignant melanoma. healthcare **2**(1), 1 (2013)
11. Soenksen, L.R., Kassis, T., Conover, S.T., Marti-Fuster, B., Gray, M.L.: Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. Science Translational Medicine **13**(581), eabb3652 (2021)
12. Tang, P., et al.: Fusionm4net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification. Medical Image Analysis **76**(102307), 1–13 (2022)
13. Tschandl, P., et al.: Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. The Lancet Oncology **20**, 938–947 (2019)
14. Tschandl, P., et al.: Human–computer collaboration for skin cancer recognition. Nature Medicine **26**, 1229–1234 (2020)
15. Yu, Z., et al.: End-to-end ugly duckling sign detection for melanoma identification with transformers. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. pp. 176–184. Springer International Publishing, Cham (2021)