RemixFormer++: A Multi-modal Transformer Model for Precision Skin Tumor Differential Diagnosis with Memory-efficient Attention

Jing Xu, Kai Huang, Lianzhen Zhong, Yuan Gao, Kai Sun, Wei Liu, Yanjie Zhou, Wenchao Guo, Yuan Guo, Yuanqiang Zou, Yuping Duan, Le Lu, *Fellow, IEEE*, Yu Wang, Xiang Chen and Shuang Zhao

***Abstract*— Diagnosing malignant skin tumors accurately at an early stage can be challenging due to ambiguous and even confusing visual characteristics displayed by various categories of skin tumors. To improve diagnosis precision, all available clinical data from multiple sources, particularly clinical images, dermoscopy images, and medical history, could be considered. Aligning with clinical practice, we propose a novel Transformer model, named RemixFormer++ that consists of a clinical image branch, a dermoscopy image branch, and a metadata branch. Given the unique characteristics inherent in clinical and dermoscopy images, specialized attention strategies are adopted for each type. Clinical images are processed through a top-down architecture, capturing both localized lesion details and global contextual information. Conversely, dermoscopy images undergo a bottom-up processing with two-level hierarchical encoders, designed to pinpoint fine-grained structural and textural features. A dedicated metadata branch seamlessly integrates non-visual information by encoding relevant patient data. Fusing features from three branches substantially boosts disease classification accuracy. RemixFormer++ demonstrates exceptional performance on four single-modality datasets (PAD-UFES-20, ISIC 2017/2018/2019). Compared with the previous best method using a public multi-modal Derm7pt dataset, we achieved an absolute 5.3% increase in averaged F1 and 1.2% in accuracy for the classification of five skin tumors. Furthermore, using a large-scale in-house dataset of 10,351 patients with the twelve most common skin tumors, our method obtained an overall classification accuracy of 92.6%. These promising results, on par or better with the performance of 191 dermatologists through a comprehensive reader study, evidently imply the potential clinical usability of our method.**

***Index Terms*— Skin Tumor Diagnosis, Multi-modality Fusion, Dermoscopy, Metadata, Differential Diagnosis**

This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFC2504700, and funded by China Postdoctoral Science Foundation under Grant 2023M743947.

Jing Xu, Lianzhen Zhong, Yu Wang, Yuan Gao, Wei Liu, Yanjie Zhou, Wenchao Guo and Le Lu are with DAMO Academy, Alibaba Group. Yuanqiang Zou is with Alibaba Cloud Computing Co. Ltd. Kai Huang, Kai Sun, Shuang Zhao and Xiang Chen are with the Department of Dermatology, Xiangya Hospital Central South University, Changsha, 410008, China. Yuan Guo and Yuping Duan are with Center for Applied Mathematics, Tianjin University.

Jing Xu, Kai Huang, Lianzhen Zhong and Yuan Gao contributed equally to this work. Shuang Zhao, Yu Wang and Xiang Chen are the corresponding authors (e-mail: shuangxy@csu.edu.cn; flimanadam@gmail.com; chenxiangck@126.com).

I. Introduction

ACCURATE and timely diagnosis of skin lesions are crucial for preventing and managing skin tumors [1]. For example, in the UK, up to 86% of melanomas can be prevented through primary and secondary prevention efforts [2]. However, this is a challenging process that requires expertise and proper equipment. Typically, skin tumor diagnosis involves visual inspection, dermoscopy imaging, and non-imaging information differentiation [3], [4]. Unfortunately, there is a lack of experienced dermatologists, making the prompt detection of skin tumors an unmet clinical need. To address this issue, the potential of artificial intelligence (AI)-based computer-aided diagnosis (CAD) [5]–[10] has been explored, showing that AI-based CAD can achieve promising performance, reducing dermatologists' workloads and benefiting a larger population. However, the application of AI-based CAD in real practice still faces obstacles, including the inadequate ability to process multi-modal data, as a dermatologist does in a realistic clinical environment (Fig. 1b). Studies have shown that using multiple data modalities in dermatology diagnosis improves the decision-making process by providing complementary information [9], [11]–[13]. There are growing interests in designing multi-modal deep learning systems to exploit this synergy. However, effective cross-modality fusion and missing modality problems remain challenges in designing high-performing deep networks [14], which have been partially addressed in our preliminary work [15]. Besides, as two main imaging modalities in dermatology, clinical images, and dermoscopy images are examined with different cognitive models by clinicians, due to their distinct imaging principles. A multi-modal skin tumor CAD model that leverages these distinctions and closely aligns with clinical practice has not received enough attention in previous work, which is the main focus of this paper.

Clinical images taken with digital cameras or even smartphones have proven their value in real-world practice and scenarios like teledermatology [16]. Dermatologists first visually retrieve global contextual information such as lesion location, shape, and color from clinical images and then focus on the lesion area to examine detailed morphological information (Fig. 1). This top-down strategy effectively relieves the limitations of human working memory [17] and is also applicable to

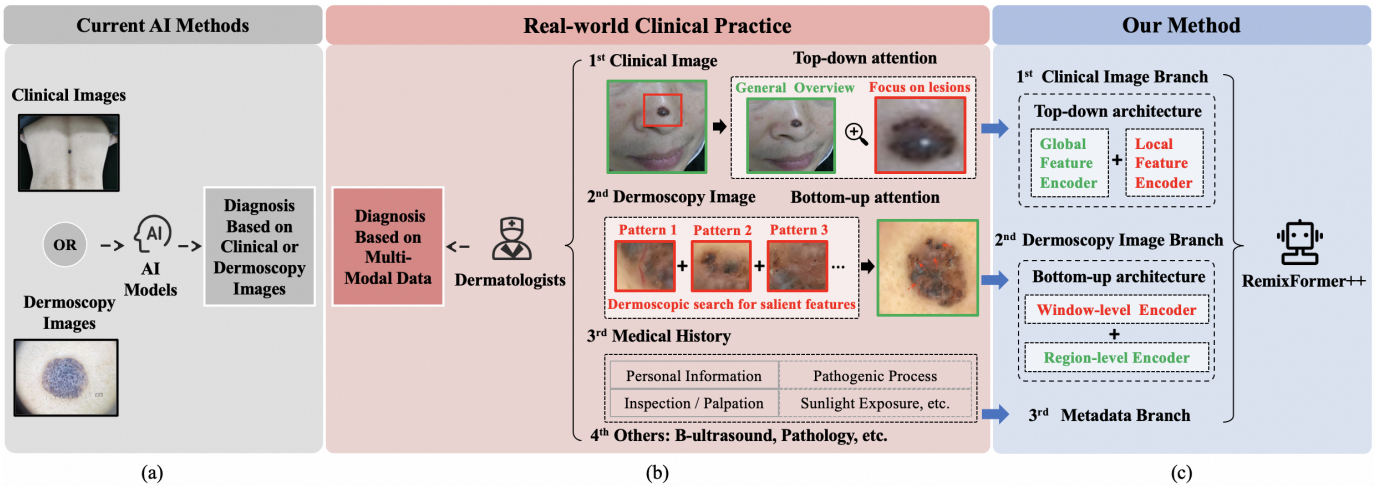


Fig. 1: (a) Skin tumor diagnosis methods only relying on single-modality information are not consistent with real clinical practices. (b) Dermatologists typically begin a general medical examination by searching anomaly via top-down attention and focusing on localized skin lesions. The patient's medical history will then be obtained. If necessary, dermoscopic screening is performed on the lesion region to depict specific patterns, such as arborized vessels, blue-gray ovoid nests, etc. Dermatologists will combine all the relevant information to make a final diagnosis. In the above case, despite the clinical image resembling a nevus, the final diagnosis is basal cell carcinoma by taking dermoscopy into account. (c) RemixFormer++ comprises three branches, effectively processing imaging and non-imaging data aligned with dermatological diagnostic procedures.

the reduction of the GPU memory footprints in AI. However, it is non-trivial for neural networks to follow this principle, as the discrete nature of the zoom-in decision makes the process non-differentiable. To model this process, the clinical image branch of our proposed RemixFormer++ framework includes a lesion selection module that uses attention sampling to overcome the non-differentiability. In addition, we include a cross-scale fusion module to effectively combine the features from low-resolution global images and high-resolution lesion patches.

With specialized optical illuminating and magnifying systems, a dermatoscope can visualize anatomical structures underneath the skin surface. To deal with high-resolution dermoscopy images, dermatologists use a memory-efficient bottom-up abstraction to grasp key features. For example, they typically scan a dermoscopy image in pieces, searching for salient textures and patterns relevant to a certain disease. Subsequently, those image pieces are conceptually represented by discovered template features for the next diagnostic reasoning; see Fig. 1b. Many deep neural networks (DNNs) have been proposed for recognizing dermoscopy images, and some of them [18], [19] attempt to incorporate important template features as supervision. However, those works are normally trained with downsampled images due to GPU memory limitation and often generate inaccurate or undesirable image embeddings with spurious features. To tackle this problem, [19], [20] have included multi-resolution crops or combining high-resolution patches, but these schemes still face limitations, as pixel-level labels are often lacking, and lesion sizes can be too large for conventional DNNs. Given these considerations, the bottom-up recognition mechanism used by dermatologists appears to be a more natural and effective way to handle high-resolution images, which is adopted in RemixFormer++. Specifically, the original image is divided into multiple windows, each of which is processed by a deep neural network to produce a high-resolution window embedding. A similar approach has been proposed in [21], but it restricts the input patch size to 224^2 to utilize pre-trained models [22]. Our RemixFormer++ differs by adopting a two-level architecture, where a self-supervised method is used to learn window-level embeddings, allowing us to choose suitable window size to balance computational cost and model performance and enabling us to utilize all available dermoscopy images, regardless of their class labels. Moreover, we design a texture attention module to automatically learn visual template features with cross-attention instead of using them as supervision.

The success of the Transformer model proves the effectiveness of the attention mechanism [23]. A complex task such as skin tumor diagnosis requires various attention mechanisms to adapt to the distinct characteristics of different data types. As discussed, for clinical images, attention is derived from contextual information in a top-down manner, and for dermoscopy images, attention is more about finding the most relevant abstract concepts with a bottom-up approach. Tailored for different modalities, these specialized and memory-efficient attention strategies are carefully implemented into our novel RemixFormer++ system for skin tumor diagnosis. By utilizing dedicated branches, the new system enhances alignment between macroscopic features in clinical images and microstructural features in dermoscopic images. Our main contributions are summarized as follows:

- For the processing channel of clinical images, the top-down attention is implemented to fuse global and local features in RemixFormer++ with several novel designs: (a) a modified Swin-Transformer encoder for global features extraction and attention computation; (b) a differentiable lesion selection module generating lesion proposals and high-resolution lesion features; (c) a cross-scale fu-

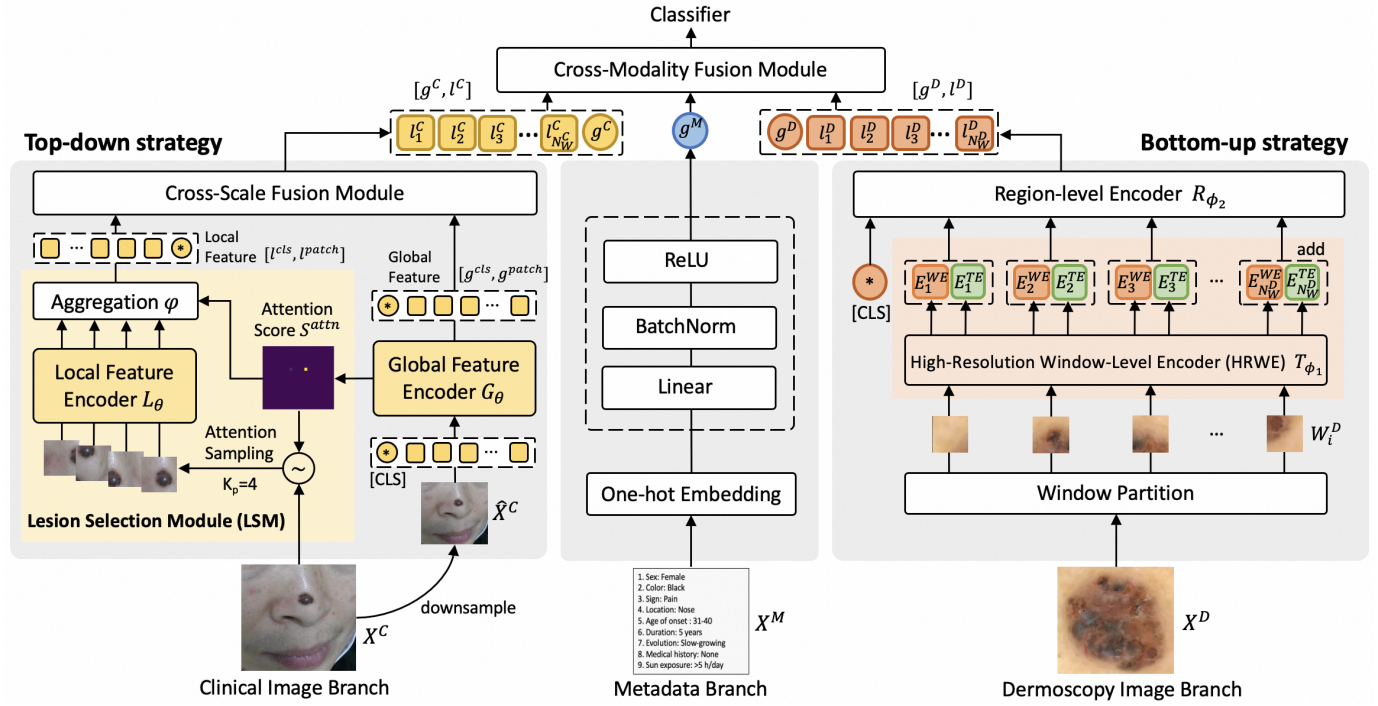


Fig. 2: The pipeline of RemixFormer++, consists of three branches: clinical image branch, dermoscopy image branch, and metadata branch. The clinical image branch has four main components: a global feature encoder generating global features and attention maps, a lesion selection module locating lesions, a local feature encoder extracting lesion features from high-resolution image patches, and a cross-scale fusion module fusing global and lesion features. The dermoscopy image branch consists of the two-level hierarchical encoder at the window and region level to extract texture information and learn high-resolution image representations. The metadata branch processes patient metadata using one-hot representation with linear embedding.

sion module fusing global and lesion features.

- For dermoscopy images, the bottom-up attention is realized via a two-level hierarchical architecture with a high-resolution window-level encoder and region-level encoder: (a) high-resolution window embeddings are learned through self-supervision, capturing detailed submicroscopic structural features; (b) texture embeddings are learned via a novel multi-scale texture attention module, serving as feature templates; (c) region-level encoder learns global context and models the dependencies among all windows.
- Our proposed RemixFormer++ demonstrates exceptional classification performance on four single-modality datasets (PAD-UFES-20, ISIC 2017, ISIC 2018, and ISIC 2019), and achieves a new benchmark accuracy of 82.5% on the publicly available multi-modal dataset Derm7pt. Our method has also been validated on a comprehensive in-house X-SkinTumor-12 dataset, gaining an accuracy of 92.6% and an F1 score of 82.4%. Moreover, in a multi-reader user study with 191 dermatologists, our method exceeds the performance of most clinical experts, indicating its strong generalizability towards promising clinical applications.

II. RELATED WORK

In recent years, advanced deep learning methods have shown great progress in many areas [20], [24] and have been applied to skin tumor recognition, among which several closely related work ranging from single-modality to multi-modality will be introduced and discussed.

A. Single-modality Methods for Skin Tumor Recognition

Esteva *et al.* [5] demonstrated the effectiveness of DNNs for general skin lesion diagnosis, which can classify malignant and benign skin tumors with a level of competence comparable to that of dermatologists. Likewise, several dermatology-related studies [7], [10], [25] have shown that deep learning-based methods in experimental settings are equivalent to or even better than human experts in skin lesion classification.

Clinical images, as one of the major skin imaging modalities, taken either by a standard digital camera or handphone, which normally presents more variations in view, angle, resolution, and lighting, prove to be valuable in clinical and teledermatology diagnosis [16]. Yang et al. [18] constructed a clinically oriented diagnostic system capable of identifying skin diseases from clinical images based on six medical representations of skin lesions according to dermatological criteria. To differentiate non-suspicious and suspicious pigmented lesions (SPLs), Soenksen et al. [8] developed a DNN-based system for SPLs detection with ugly duckling identification in wide-field dermatological photographs. Wu *et al.* [26] investigated skin tumor classification using the Xiangya–Derm clinical skin dataset and five mainstream convolutional neural networks, exploring various algorithms for this purpose. Zhao *et al.* [27] employed the Xception architecture to build a

risk-level classifier, competing against and outperforming 20 professional dermatologists.

Dermoscopy images allow for clearer observation of skin lesions and enable dermatologists to analyze specific morphological features for skin tumor differential diagnosis. With the development of the largest public dermoscopy image dataset by the International Skin Imaging Collaboration (ISIC) [25], several diagnostic algorithms have made significant progress in skin tumor recognition based on dermoscopy data. Xie *et al.* [28] proposed a mutual bootstrapping deep convolutional neural network model for simultaneous skin lesion segmentation and classification. Khan *et al.* [29] proposed a two-stream deep neural network information fusion framework for multiclass skin cancer classification. To simulate doctors' diagnosis process, [30] proposes a clinical-inspired network that includes a lesion area attention module, a feature extraction module, and a lesion feature attention module. Although these methods have demonstrated advanced performance, they exhibit shortcomings in handling the complexity of texture and processing high-resolution images.

B. Multi-modality Methods for Skin Tumor Recognition

As multi-modality databases of skin tumors continue to expand, research is shifting from relying on single-modality approaches to integrating multiple modalities. Pacheco *et al.* [12] constructed a skin tumor dataset (PAD-UFES-20) comprising patient data and clinical images collected from smartphones. They employed a DNN-based approach with an aggregation mechanism to fuse this two-modality information, which can improve the accuracy of diagnosis. Similarly, by adding metadata such as age, anatomical site, and gender to a dermoscopy model, Gessert *et al.* [31] verified that performance was improved by ensembles of EfficientNet [32].

Incorporating clinical images and dermoscopy images, Ge *et al.* [6] proposed a multi-modality DNN-based architecture to extract discriminative features from two image modalities and showed the effectiveness of a multi-modality approach for skin tumor classification. Furthermore, Haenssle *et al.* [7] found that AI models generally outperformed most dermatologists when solely relying on dermoscopy images. However, when multi-modal information was provided, the performance of dermatologists was on par with that of AI. Thus, it is reasonable to assume that neural networks can achieve better performance when trained with multi-modal data as dermatologists do.

Most recently, Fu *et al.* [33] proposed a graph-based inter-category and intermodality network, encoding the relationship between diagnoses and categories on Derm7pt [11]. With three-modality data, Tang *et al.* [9] constructed a two-stage method that concatenated feature information from clinical and dermoscopy images, and then combined metadata via SVM-based clustering. Our preliminary work [15] presented an efficient multi-modal transformer-based model, which included a novel sampling strategy and an effective cross-modality fusion module (CMF), confirming the performance-critical importance of utilizing multi-modal data for more accurate skin tumor diagnosis.

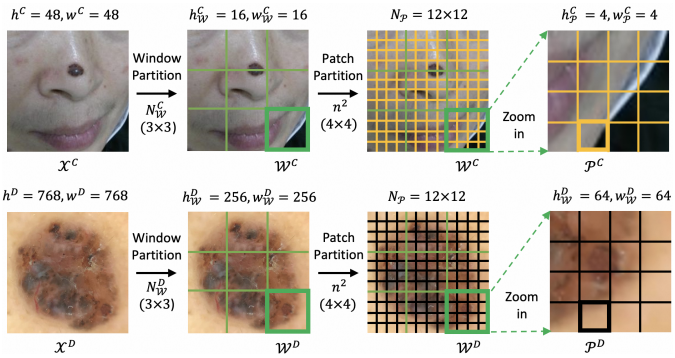


Fig. 3: Illustration of the original image, window image, and patch image, where $\mathcal{X}^C$ and $\mathcal{X}^D$ denote the original clinical image and dermoscopy image, $\mathcal{W}^C$ and $\mathcal{W}^D$ denote the window-sized clinical image and dermoscopy image, $\mathcal{P}^C$ and $\mathcal{P}^D$ denote the patch-sized clinical image and dermoscopy image. The image resolutions shown (48^2 or 768^2) are for illustration purposes only and do not represent the actual size of the images used in the experiments.

III. METHODS

A. Overview

Closely following clinical practices, RemixFormer++ consists of three branches: clinical image branch, dermoscopy image branch, and metadata branch. Fig. 2 illustrates the overall framework of our model, where both image features and metadata embeddings are fed into the cross-modality fusion module to form a global representation for final disease classification. The details of each branch are described below.

B. Clinical Image Branch

The diagnostic value of clinical images has received great attention [34]. In the typical inspection of clinical images, dermatologists take into consideration both global information and local information. To mirror the real-world examination process, we use a global encoder to capture features such as color, shape, and body parts; whereas using a local encoder to zoom in on detailed attributes such as macule, plaque, scale, etc. We denote $\mathcal{X}^C$ as the clinical image with size $h^C \times w^C$, which is divided by multiple windows $\mathcal{W}_i^C$, and each window is formed by several patches $\mathcal{P}_j^C$, as shown in Fig. 3. In our clinical image branch, the global and local feature encoders (G_θ and L_θ) are two networks based on the Swin-Transformer block, where θ denotes its parameters. Note that the high-resolution clinical image $\mathcal{X}^C$ is down-sampled to the $\widehat{\mathcal{X}}^C$ before input into G_θ to reduce memory consumption. To guarantee the local encoder captures fine-grained details in an end-to-end manner, we introduce a Lesion Selection Module (LSM) to select lesion areas from the original high-resolution image $\mathcal{X}^C$ other than down-sampled images $\widehat{\mathcal{X}}^C$ as [35], or using a two-stage approach [36]. To this purpose, an additional class token is inserted into the standard Swin-Transformer block to facilitate the attention computation.

1) Global Feature Encoder: Due to its effectiveness and low complexity, Swin-Transformer [37] is adopted as the backbone of both global and local encoders. Normally, the learnable "classification token" known as the class token or [CLS] token

is inserted into the sequence of embedded patches to aggregate information from ViTs. The classification is performed by a multilayer perceptron (MLP) head that processes the [CLS] token from the last layer. In contrast, without the [CLS] token, Swin Transformers perform classification by applying a global average pooling layer to the feature map of the last stage, followed by a linear classifier. The handling in the Swin Transformer can be conceptualized as a virtual class token that aggregates information from patch tokens through average pooling. The patch tokens that exhibit strong correlations with the class token, regardless of whether it is physically present or virtual, are indicative of patches carrying crucial lesion information. Calculating these correlations involves analyzing the attention values between the patch tokens and the class token. While straightforward for ViTs, special handling is necessary for Swin Transformers due to their unique architecture. To facilitate this process, we integrate a class token into the Swin Transformer architecture. We modify the shifted window-based multi-head self-attention (WMSA) module in a normal Swin-Transformer block to incorporate a class token [CLS]. Following the Swin-Transformer, the patch image $\mathcal{P}_j^C$, $j = 1, \ldots, N_{\mathcal{P}}^C$, of size $h_{\mathcal{P}}^C \times w_{\mathcal{P}}^C$ (4×4 in our implementation) is firstly transferred to the patch token, and the self-attention is computed from patch tokens within a window. The number of tokens along the vertical axis and the horizontal axis in a window is set to be equal, e.g. $n \times n$ (7×7 in our implementation). Thus, a window image $\mathcal{W}_i^C$ represents an image cropping with the size of $h_{\mathcal{W}}^C = n \times h_{\mathcal{P}}^C, w_{\mathcal{W}}^C = n \times w_{\mathcal{P}}^C$, Fig. 3 illustrating an example. Let $x \in R^{(N_{\mathcal{P}}^C+1)\times d}$ be the input of the MSA module formed by $N_{\mathcal{P}}^C$ patch tokens and [CLS] token, where d is the number of channels and $N_{\mathcal{P}}^C = (h^C/h_{\mathcal{P}}^C) \times (w^C/w_{\mathcal{P}}^C)$. The self-attention of the [CLS] token can be computed by

$$z^{cls} = \text{SoftMax}\left(\frac{q_0 \cdot K^T}{\sqrt{d/H}}\right) \cdot V \in R^{1\times d}, \tag{1}$$

where q_0 is the [CLS] token as well as the first token, i.e. $q_0 = Q[0,:] \in R^{1\times d}$, H is the number of heads and the [CLS] token is assumed to be the first token. The corresponding $query$, key, and $value$ embedding are denoted by $Q, K, V \in R^{(N_{\mathcal{P}}^C+1)\times d}$. To compute the self-attention of patch tokens within local windows, we perform a window partitioning (WP) scheme on patch tokens, taking Q as an example

$$Q^{WP} := \text{WP}(Q) = [q_1, q_2, \ldots, q_{N_{\mathcal{W}}^C}] \in R^{N_{\mathcal{W}}^C \times (n^2+1)\times d}, \tag{2}$$

where q_i denotes the patch tokens within the window, i.e. $q_i \in R^{(n^2+1)\times d}$ for $i \in \{1, 2, \ldots, N_{\mathcal{W}}^C\}$, $N_{\mathcal{W}}^C = N_{\mathcal{P}}^C/n^2$, and n^2 is the number of patch tokens in a window. Note that the first token in each q_i corresponds to the $query$ of the [CLS] token. Similarly, we can define $K^{WP} = [k_1, k_2, \ldots, k_I]$ and $V^{WP} = [v_1, v_2, \ldots, v_I]$, where the first tokens in each k_i and v_i correspond to the key and $value$ of the [CLS] token, respectively. Then the self-attention of the patch tokens within a window becomes

$$z_i = \text{SoftMax}\left(\frac{\bar{q}_i \cdot k_i^T}{\sqrt{d/H}}\right) \cdot v_i \in R^{n^2\times d}, \tag{3}$$

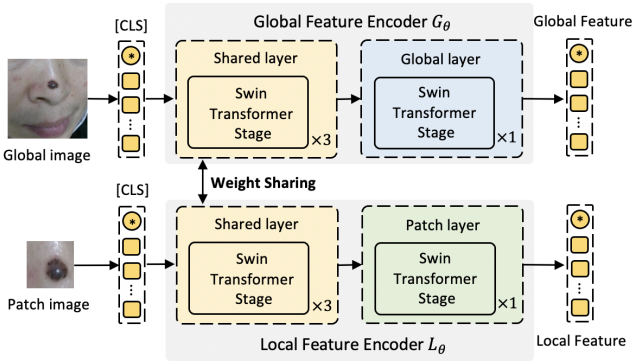


Fig. 4: The composition of the global encoder G_θ and local encoder L_θ in our clinical image branch.

where $\bar{q}_i = q_i[1 : n^2, :] \in R^{n^2\times d}$. All patch tokens z^{patch} can be obtained via the reverse window partition

$$z^{patch} = \text{reverse\_WP}([z_1, z_2, \ldots, z_{N_{\mathcal{W}}^C}]) \in R^{N_{\mathcal{P}}^C \times d}. \tag{4}$$

The above computation is implemented in a module named window-based multi-head self-attention with a class token (shorted by CLS_WMSA), and the corresponding output for a given x is defined as

$$z = CLS\_WMSA(x) = \left[z^{cls}, z^{patch}\right] \in R^{(N_{\mathcal{P}}^C+1)\times d}. \tag{5}$$

The original WMSA modules in Swin-Transformer are replaced by CLS_WMSA, and the global encoder G_θ is formed by stacking four stages of the modified Swin-Transformer blocks, as demonstrated in the upper part of Fig. 4. The global features produced by the last stage of G_θ denoted as $[g^{cls}, g^{patch}]$ are fused with the corresponding local features from the local encoder for further classification.

2) Lesion Selection Module (LSM): With the above modifications, the [CLS] token communicates with each patch token via self-attention operation in each stage of the G_θ and is further used for disease classification after feature fusion. The patch token strongly correlated to the [CLS] token indicates the corresponding patch has a high probability of containing a lesion. Generally speaking, the high-level features have more semantic information, similar to [38]. Thus, the [CLS] token from the last stage is used to compute attention scores S^{attn} as follows

$$O_1^h = Q_0^h \cdot K_{h,1:}^T \in R^{1\times N_{\mathcal{P}}}, \qquad h = 1, 2, \cdots, H, \tag{6}$$

$$O_2^h = \left(Q_{1:}^h \cdot {K_{h,0}}^T\right)^T \in R^{1\times N_{\mathcal{P}}}, \quad h = 1, 2, \cdots, H, \tag{7}$$

$$S^{attn} = \frac{1}{H}\sum_{h=1}^{H} \text{SoftMax}\left(O_1^h\right) \odot \text{SoftMax}\left(O_2^h\right), \tag{8}$$

where Q^h and K^h are the $query$ and key of each attention head in a CLS_WMSA module, Q_0^h and K_0^h are the $query$ and key of [CLS] token in each attention head, and $Q_{1:}^h$ and $K_{1:}^h$ are the $query$ and key of patch tokens.

The most salient image patches are extracted from raw image $\mathcal{X}^C$, denoted by $\{\mathcal{P}_{j_k}^C\}_{k=1}^{K_p}$, where patches with K_p ($K_p = 4$ in our implementation) largest S^{attn} values are chosen to be the input of the local feature encoder L_θ. Since

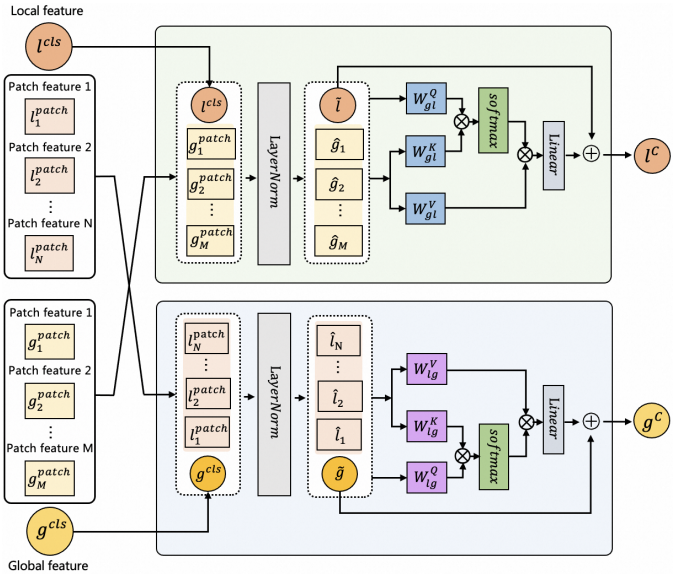


Fig. 5: The cross-scale fusion module in our clinical image branch.

S^{attn} is computed on the down-sampled feature map, the actual size of the lesion patches is enlarged by a fixed ratio to guarantee that the lesion area is covered. Then the local features $[l^{cls}, l^{patch}] \in R^{K_p \times (N_L+1) \times d}$ can be obtained, where N_L denotes the number of local patch tokens.

When the lesion selection module (LSM) is implemented with the top-K operation, the non-differentiability will break the gradient flow from the attention error. Therefore, inaccurate initial attention can stick the network at irrelevant image regions, which makes end-to-end training difficult. We adopt the attention sampling technique in [39] to relax the non-differentiability by a Monte Carlo approximation of the expectation with the sampling subjected to attention distribution. Specifically, we sample a set V of N i.i.d. indices from the attention distribution, i.e., $V = \{v_i \sim S^{attn} | i \in \{1, 2, \cdots, N\}\}$ and estimate the aggregated local features φ as follows

$$\varphi = \frac{1}{N} \sum_{v \in V} L_\theta([l^{cls}, l^{patch}]). \tag{9}$$

3) Local Feature Encoder: The local encoder L_θ is also built with four stages of the modified Swin-Transformer blocks. To save model parameters, the first three stages share weights with the G_θ, and the last stage is independent, as shown in Fig. 4. To optimize GPU memory usage and retain maximum detailed information while excluding context information already embedded in global features, we employ LSM to selectively extract lesion patches from the original high-resolution image. Instead of directly using $\mathcal{X}^C$, the selected lesion patches are taken as input for the local encoder. The local features $[l^{cls}, l^{patch}]$ produced from the last stage of L_θ, together with the global features are then fed into the cross-scale fusion module for feature alignment.

4) Cross-scale Fusion Module: Our previous work [15] introduced a cross-modality fusion module designed to integrate multiple sources of data. It can effectively align and combine features from various inputs to enhance the common semantic information existing in the different features. We implement the same strategy to integrate local features and global features, which is named by the cross-scale fusion (CSF) module. As depicted in Fig. 5, the local features $[l^{cls}, l^{patch}]$ and global features $[g^{cls}, g^{patch}]$ are firstly exchanged and assembled to new tensors. After a Layer Normalization, the new global and local features $\hat{l}, \tilde{g}$ and $\hat{g}, \tilde{l}$ are created and used to calculate the $query$, key, and $value$ embedding as

$$\mathcal{Q}^g = \tilde{\boldsymbol{g}} \boldsymbol{W}_{lg}^{\mathcal{Q}}, \quad \mathcal{K}^g = \hat{\boldsymbol{l}} \boldsymbol{W}_{lg}^{\mathcal{K}}, \quad \mathcal{V}^g = \hat{\boldsymbol{l}} \boldsymbol{W}_{lg}^{\mathcal{V}}, \tag{10}$$

$$\mathcal{Q}^l = \tilde{\boldsymbol{l}} \boldsymbol{W}_{gl}^{\mathcal{Q}}, \quad \mathcal{K}^l = \hat{\boldsymbol{g}} \boldsymbol{W}_{gl}^{\mathcal{K}}, \quad \mathcal{V}^l = \hat{\boldsymbol{g}} \boldsymbol{W}_{gl}^{\mathcal{V}}. \tag{11}$$

Then we compute the cross-attention by

$$\boldsymbol{M}_{att}^g = \text{softmax}(\frac{\mathcal{Q}^g (\mathcal{K}^g)^T}{\sqrt{F/h}}), \quad \boldsymbol{M}_{cross}^g = \boldsymbol{M}_{att} \mathcal{V}^g, \tag{12}$$

$$\boldsymbol{M}_{att}^l = \text{softmax}(\frac{\mathcal{Q}^l (\mathcal{K}^l)^T}{\sqrt{F/h}}), \quad \boldsymbol{M}_{cross}^l = \boldsymbol{M}_{att} \mathcal{V}^l. \tag{13}$$

With exchanging information between global and local features through multi-head attention in CSF, the final features $[g^C, l^C]$ of the clinical image $\mathcal{X}^C$ are obtained as follows

$$\boldsymbol{g}^C = \tilde{\boldsymbol{g}} + \text{linear}(\boldsymbol{M}_{cross}^g), \quad \boldsymbol{l}^C = \tilde{\boldsymbol{l}} + \text{linear}(\boldsymbol{M}_{cross}^l). \tag{14}$$

C. Dermoscopy Image Branch

As discussed in Related Work, many works have delved into dermoscopy images [25], [28]–[30]. Despite their effectiveness, these methods struggle with the complex details in dermoscopic images, like fine textures in high-resolution images. Due to the high computation cost and memory consumption, they require downsizing images and inevitably cause the loss of details. Inspecting high-resolution dermoscopy images is also a visual task with a heavy burden for dermatologists. To simplify the inspection process, dermatologists first look for some inter-media features such as globules, regression areas, and streaks from high-resolution patches, and assemble these salient semantic features for the final reasoning. Inspired by the information reduction in this process, we introduce a two-level hierarchical architecture to learn representations from high-resolution images. Firstly, the high-resolution window-level encoder (HRWE), a ViT [40] equipped with novel multi-scale texture attention (MSTA), is used to learn window embeddings and texture embeddings from high-resolution windows. The second stage region-level encoder takes the non-overlapping window-level embeddings as input and learns global context and models the dependencies among all windows. A proper combination of the window size and the embedding dimension becomes an important choice for balancing training efficiency and embedding's representativeness.

1) High-Resolution Window-Level Encoder (HRWE): A high-resolution dermoscopy image $\mathcal{X}^D$ is split into non-overlapping window images $\mathcal{W}_i^D$ with size $h_{\mathcal{W}}^D \times w_{\mathcal{W}}^D$ ($h_{\mathcal{W}}^D = w_{\mathcal{W}}^D = 256$ in our implementation), where $i \in \{1, 2, \ldots, N_{\mathcal{W}}^D\}$ and $N_{\mathcal{W}}^D$ the number of windows. Since these windows are large enough to contain rich semantic information, we feed them into a self-supervised pre-trained ViT model denoted by T_{ϕ_1} with ϕ_1 being the parameters, which works as the encoder to

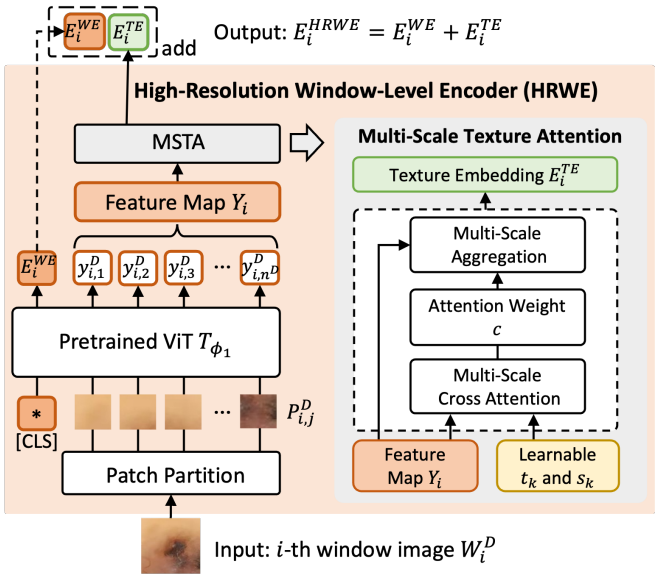


Fig. 6: The pipeline of the high-resolution window-level encoder.

replace simple linear projection; see Fig. 6. The learnable class token [CLS] at the output of the T_{ϕ_1} serves as the window embedding E^{WE} and an MSTA module is proposed to capture texture information E^{TE}.

Self-supervised pre-trained ViT for window embedding: Supervised learning is not the most appropriate paradigm to train T_{ϕ_1}, as the exact labels are not available for window images and the labeling process is laborious and subjective. Therefore, we adopt the self-distillation approach DINO in [41] to train T_{ϕ_1} with self-supervised learning (SSL) on massive 256^2 window images. More specifically, for each $\mathcal{W}_i^D$, we employ a multi-crop strategy to generate a set of diverse views including two global views encompassing the entire image, and multiple local views with smaller resolutions. All crops including both global and local views are processed by the student network, while only the global views are passed through the teacher network. The teacher and student networks have identical architectures but different parameter values. The teacher's parameters are updated by computing an exponential moving average of the student's parameters over time. At convergence, the final teacher network is used as the pre-trained T_{ϕ_1}.

The pre-trained T_{ϕ_1} is further trained with the whole dermoscopy branch via end-to-end supervised learning. Similar to normal ViTs, during the training, the i-th window image $\mathcal{W}_i^D$ is divided into patches $\mathcal{P}_{i,j}^D$, where for $j \in \{1, 2, \ldots, n^D\}$ with n^D being the total number of patches in a window, which are linearly projected into tokens. After prepending a class token [CLS] and adding position embeddings, all tokens pass through T_{ϕ_1}. The [CLS] token from the last layer of the T_{ϕ_1}, denoted by E_i^{WE}, is later used for the classification. The remaining output of the last layer Y_i is then fed into the multi-scale texture attention module. Specifically, $Y_i \in R^{n^D \times d_s}$ with d_s denoting the dimension,

$$Y_i := [y_{i,1}, y_{i,2}, \ldots, y_{i,n^D}] = T_{\phi_1}(\mathcal{P}_{i,1}^D, \mathcal{P}_{i,2}^D, \ldots, \mathcal{P}_{i,n^D}^D),$$

where $y_{i,j}$ denotes the j-th patch embedding in the i-th window.

Multi-Scale Texture Attention (MSTA) for texture embedding: Inspired by [42], we designed a new texture attention module to learn feature templates with more efficient and scalable cross-attention. We feed Y_i into MSTA to learn texture information. Specifically, $\{t_k \in R^{d_s}\}_{k=1}^{K_s}$ is a set of vectors ($K_s = 32$) resembling templated features stored in the dermatologists' long-term memory, and $\{s_k \in R^{d_s}\}_{k=1}^{K_s}$ is defined as a set of learnable scaling factors accounting for the multi-scale nature of texture embedding. Inside a given window, the patch embedding and texture templates are defined as

$$a_{j,k} = y_{i,j} \cdot (t_k)^T \quad \text{and} \quad c_{j,k} = \frac{exp(-s_k a_{j,k})}{\sum_{\ell=1}^{K_s} exp(-s_\ell a_{j,\ell})}. \quad (15)$$

We directly use dot-product attention to compute $a_{j,k}$, instead of using the residuals as in [42]. Then, the global vectors t_k can gather the information from each patch embedding in all windows and serve as texture templates. We also use a learnable s_k to account for the scaling effects for different texture templates and obtain the aggregation weights by scaled attention. Then the learnable parameters t_k and s_k are trained via supervised learning.

The texture attention module essentially calculates the resemblance between patch embedding and texture templates and assembles them accordingly to texture embedding as follows

$$E_{i,k}^{TE} = \sum_{j=1}^{n^D} c_{j,k} y_{i,j}. \quad (16)$$

The texture embedding E_i^{TE} is formed by concatenating all $E_{i,k}^{TE}$ and passing through a linear layer. The final embedding of i-th window $\mathcal{W}_i^D$ can be expressed as

$$E_i^{HRWE} = E_i^{WE} + E_i^{TE}. \quad (17)$$

2) Region-Level Encoder: After prepending an image [CLS] token E^{cls} to all window embeddings E^{HRWE}, we serve the concatenated sequence $E = [E^{cls}, E^{HRWE}] \in R^{(N_{\mathcal{W}}^D+1) \times d_s}$ as input to the region-level encoder R_{ϕ_2} composed of a sequence of transformer blocks, aiming to learn global context between windows. The output feature map of the last layer in R_{ϕ_2} is defined as $[g^D, l^D] \in R^{(N_{\mathcal{W}}^D+1) \times d_s}$, where the final representations of g^D and $l^D = \{l_i^D\}_{i=1}^{N_{\mathcal{W}}^D}$ denote the global feature and local feature of the image $\mathcal{X}^D$, respectively. It is challenging to make the whole training process converge without a pre-trained T_{ϕ_1}.

D. Metadata Branch

As the metadata used in our study are all structured data, we can use simple one-hot embedding instead of complex text encoders to represent metadata. Normally, metadata has N^M categories, and for each category, there are d_i options, where $i \in \{1, 2, \ldots, N^M\}$. Using one-hot encoding, each category can be represented by a d_i-sized vector, with a single entry valued at one and the remaining entries equal to zero. The metadata $\mathcal{X}^M$ for each sample is formed by

concatenating these vectors, and its length is $\sum_i^{N^M} d_i$. Given this representation, the $\mathcal{X}^M$ is highly sparse and lacks semantic information. However, it is adequate for most existing skin tumor datasets, including our private one. Subsequently, the metadata feature g^M is obtained by processing $\mathcal{X}^M$ with metadata branch M_θ, which consists of a linear layer followed by batch normalization and ReLU activation, e.g. $g^M = M_\theta(\mathcal{X}^M)$.

Finally, the global features g^C, g^D and local features l^C, l^D of the clinical image $\mathcal{X}^C$ and dermoscopy image $\mathcal{X}^D$, and metadata feature g^M of the patient's metadata are fed into the effective CMF module to form a global representation for the final disease classification.

IV. EXPERIMENTS

A. Dataset

We employ four skin tumor datasets in this study: three public and one private dataset.

1) PAD-UFES-20 dataset: The public clinical dataset PAD-UFES-20 [12] contains 2,298 samples of 6 skin lesions, which are basal cell carcinoma (BCC), squamous cell carcinoma (SCC), actinic keratosis (AK), seborrheic keratosis (SK), melanoma (Mel), and nevus (Nev). In our comparative experiments, we adopted the 5-fold cross-validation strategy, consistent with other compared methods [12], [43].

2) ISIC dataset: Three public dermoscopy datasets ISIC 2017, ISIC 2018, and ISIC 2019 are selected for experiments. Among them, The ISIC 2017 contains 2,000 training images and 600 testing images, including 3 diseases and 4 attributes. The ISIC 2018 contains 10,015 training images and 1,512 testing images, including 7 diseases and 5 attributes. The ISIC 2019 contains 25,331 training images and 8,238 testing images, including 8 diseases. The attributes include milia-like cysts (MLCs), negative pigment network (NPN), pigment network (PN), streaks (STR), and globules (GLO). Due to the unavailability of open-source labels for the ISIC 2019 test set, we also applied the 5-fold cross-validation strategy to the training set, in alignment with the comparative method CI-Net [30].

3) Derm7pt dataset: The publicly available multi-modal Derm7pt [11] dataset contains 413 training cases, 203 validation cases, and 395 testing cases. Each case comprises a dermoscopy image, a clinical image, and 7-point checklist labels, the diagnostic (DIAG) label is divided into 5 types: BCC, Nev, Mel, SK, and miscellaneous (Misc). The 7-point checklist labels are PN, blue whitish veil (BWV), vascular structures (VS), pigmentation (PIG), STR, dots and globules (DaG), and regression structures (RS). Each category has different types, including absent (ABS), present (PRS), typical (TYP), atypical (ATP), regular (REG), and irregular (IR). We employed the same data splits as the original paper [11] for all experiments.

4) X-SkinTumor-12 dataset: Our private dataset named X-SkinTumor-12 was collected from Xiangya Hospital from 2016 to 2022, and annotated by dermatologists with at least six years of experience. The dataset contains 29,263 images and 10,351 patients. Each patient contains clinical images (C), dermoscopy images (D), and a set of metadata (M). The detailed data distribution is shown in Table I. The patients' metadata with nine attributes is shown in Table II. The dataset has 12 types of skin tumors: SCC, Mel, BCC, Bowen's disease, Paget's disease, AK, keloid (Kel), dermatofibroma (DF), sebaceous nevus (SN), SK, Nev and haemangioma (Hem). All experiments was performed with 5-fold cross-validation on the large-scale X-SkinTumor-12 for consistency.

TABLE I: The detailed statistics for X-SkinTumor-12. #Total represents the total number of patients in each category, where the number of three-modality paired data is shown by #Pair. #C and #D denote the number of clinical and dermoscopy images in each category, respectively.

Disease	Patient-level		Image-level	
	#Total	#Pair	#C	#D
BCC	706	357	1398	721
Mel	365	207	1040	903
SCC	371	126	611	166
Bowen	328	32	425	95
Paget	199	16	487	87
AK	316	73	506	115
Kel	185	108	299	199
DF	140	75	223	89
SN	215	145	380	243
SK	2869	2472	3468	2947
Nev	4547	4270	6920	7675
Hem	110	86	167	99

TABLE II: Description of metadata on X-SkinTumor-12.

Attribute	Description
Sex	The sex of the patient. (Options: Male, Female)
Color	Color of skin lesions. (Options: Yellow, Brown, Black, Red, Other)
Sign	Medical signs perceived by the patient. (Options: Pain, Itching, Bleeding, No hair, None)
Location	Location of skin lesions. (Options: Face, Neck, Trunk/Limbs, Genitals, Head, Breast, Other)
Age of onset	The age that the skin lesion appeared. (Options: <20, 20-30, 31-40, 41-50, >50)
Duration	Duration of skin lesions. (Options: <5 years, 5-10 years, >10 years)
Evolution	The process of skin lesions growing. (Options: Trauma, Born, Slowly growing, Rapidly growing, None)
Medical history	Past medical history. (Options: Yes, No, Unknown)
Sun exposure time	Sun exposure time of a day. (Options: <2 h/day, 2-5 h/day, >5 h/day)

B. Implementation Details

For the clinical image branch, we use Swin-T for the backbone G_θ and L_θ and experiment with ViT-S/16 and ViT-B/16. For our dermoscopy image branch, we choose the lightweight ViT-S/16 (384 token dimension, 12 layers, and 6 heads) as the backbone of T_{ϕ_1} in the HRWE module, and the transformer blocks in R_{ϕ_2} have the same configuration as those in T_{ϕ_1}. To train the HRWE module using self-supervised learning, we utilized dermoscopy images from two sources: the ISIC 2019 dataset (encompassing ISIC 2018) and our proprietary collection, X-SkinTumor-12. The ISIC

2019 dataset comprises 33,569 images, while X-SkinTumor-12 contains 13,339 dermoscopy images. We partitioned the ISIC 2019 dataset into two sets. The first set contains 10,000 randomly selected images, cropped into 256^2 windows. It is worth mentioning that the 256^2 window size provides an appropriate field of view to cover detailed submicroscopic structural features in dermoscopic images, such as pigment networks, globules, and other patterns, typically within the sub-millimeter scale [44]. This chosen size is sufficient for our model to capture the essential microstructural details for reliable analysis. The rest are resized to 1024^2 and then split into windows of 256^2. The X-SkinTumor-12 dataset underwent a similar procedure, except that the second set is scaled to 2048^2, as the images in X-SkinTumor-12 have a higher resolution. The resultant dataset, comprising a total of 2,007,437 images, was used for self-supervised learning. The data augmentation includes random flip with color jittering, Gaussian blur, polarization, and multi-crop with a bicubic interpolation, where flip probability is 0.5, and color jittering is implemented by the "Colorjitter" function in pytorch with brightness=0.4, contrast=0.4, saturation=0.2 and hue=0.1. Similar to the standard multi-crop setup, we utilized two global views at a resolution of 256^2 and employed multiple local views with a resolution of 96^2 to capture smaller regions (e.g., less than 50%) of the original image. The temperature for the student network is set to 0.1, while a linear warm-up for the temperature of the teacher network is used, starting from 0.04 to 0.07 in the first 30 epochs. The batch size is set to be 256 for the training process.

We perform comparative experiments on different data modalities. More specifically, all raw clinical images in PAD-UFES-20 and Derm7pt are resized to 896^2 and downsampled again to 384^2 as input to G_θ. The LSM module selects 224^2 patches, which corresponds to several 3×3 areas on the feature map of the last layer in G_θ, from the 896^2 input and feeds them into L_θ. Dermoscopy images are resized to 1024^2 and then divided into 256^2 patches, fed to the HRWE module, for all ISIC datasets and Derm7pt. For X-SkinTumor-12, clinical images are also resized to 896^2 but all dermoscopy images are set to 2048^2 to match the average image size of the dataset. We augment the training data with a pipeline of transformations to account for geometric and color variations. This pipeline includes: vertical and horizontal flips (50% probability each); random crop with area scaling between 0.65 and 1.0 of the original image; random rotation with a degree uniformly sampled between 0 and 180 (in multiples of 10); random affine transformation with translations up to 15% of the image height and width, and shears between -10 and 10 degrees; and brightness, contrast, and saturation jittering with values ranging from 0.5 to 1.5. We do not apply any augmentation on validation and test data except for image resizing. All models are trained for 200 epochs on 4 NVIDIA Tesla V100 GPUs with batch size 64. We employ an SGD optimizer with a cosine learning rate schedule, and the initial learning rate is 1e-4. The model parameters are initialized by ImageNet pre-trained weights to speed up the convergence, except that the T_{ϕ_1} uses self-supervised pre-trained weights. All networks are trained in an end-to-end manner using a cross-entropy loss.

C. Comparison Methods

During our numerical experiments, we compare the proposed RemixFormer++ with several state-of-the-art multi-modality methods on skin tumor datasets, which are described as follows.

- Triple-Net [6]: The Triple-Net developed a deep convolutional neural network architecture with saliency feature descriptors to capture discriminative features from dermoscopy and clinical images.
- EmbNet [45]: The EmbNet presented a convolutional neural network for automatic skin lesion diagnosis by combining multiple imaging modalities together with patient metadata.
- Incep-co [11]: The Incep-co proposed a multitask deep convolutional neural network to classify the 7-point melanoma checklist criteria and perform lesion diagnosis, which was trained on multimodal data including clinical and dermoscopic images, and patient metadata.
- HcCNN [13]: The HcCNN proposed a convolutional neural network with an additional hyper-branch integrating intermediary image features hierarchically, to learn more complex combinations from all stages of the network.
- FM4Net [9]: The FM4Net proposed a two-stage multi-modal learning method for multi-label skin lesion classification, which utilized the clinical and dermoscopy images, and patient's metadata information in the first stage and second stage, respectively.
- GIIN [33]: The GIIN incorporated the graph-based relationship modeling module to capture the intercategorical relationships among different attributes.
- AMFAM [46]: The AMFAM incorporated an attention-based reconstruction module to constrain the CNN backbone, ensuring that the feature representations of each modality are more discriminative.
- Remix [15]: The Remix proposed a transformer model for skin tumor differential diagnosis, where a disease-wise pairing-based remix operation and cross-modality fusion were introduced for multi-modality data learning.
- CI-Net [30]: The CI-Net used lesion area attention, feature extraction, lesion feature attention, and distinguish modules to simulate the zooming, observing, and comparing steps in the diagnostic process.

D. Experimental Setup

We first compare the performance of the clinical image branch and the dermoscopy image branch with state-of-the-art methods on single-modality datasets (PAD-UFES-20, ISIC 2017, ISIC 2018, and ISIC 2019). Then we benchmark RemixFormer++ against existing multi-modal methods on Derm7pt. In addition, we perform extensive ablation studies to evaluate the effectiveness of LSM, CSF, and HRWE on X-SkinTumor-12. To demonstrate the clinical usability of our method, we conducted a reader study to compare the performance of RemixFormer++ against 191 dermatologists in a simulated clinical setting. We use the area under the curve (AUC), macro-averaged F1, sensitivity (Sens), precision (Prec), specificity

TABLE III: Comparison of our clinical image branch with other methods on PAD-UFES-20 dataset. *BACC is the gold metric (%).

Modality	Method	Acc	AUC	*BACC
C	Pacheco et al. [12]	70.7±2.8	**93.2±1.0**	65.8±4.5
	Ou et al. [43]	61.6±5.1	90.1±0.7	65.1±5.0
	ViT-B/16 [40]	70.6±1.6	89.6±1.1	74.5±2.3
	Swin-T [37]	71.2±3.1	89.7±1.4	75.8±2.7
	Ours-C	**72.6±2.5**	90.1±1.1	**76.3±2.2**
CM	Pacheco et al. [12]	78.8±2.5	**95.8±0.7**	75.0±3.3
	Ou et al. [43]	76.8±2.2	94.7±0.7	77.5±2.2
	ViT-B/16 [40]	76.8±1.3	92.2±1.2	79.3±1.6
	Swin-T [37]	78.8±2.5	92.9±0.7	80.9±3.1
	Lima et al. [47]	N.A.	94.1±0.6	80.0±0.6
	Ours-CM	**79.8±1.5**	93.5±1.2	**81.3±2.5**

(Spec), overall accuracy (Acc), balanced accuracy (BACC), and average precision (AP) as the evaluation metrics.

V. Results

A. Comparison with SOTA methods on PAD-UFES-20 dataset

We compare the clinical image branch with other methods on PAD-UFES-20 via 5-fold cross-validation. As listed in Table III, when only clinical images are used, our model significantly outperforms the CNN-based methods reported in [12], [43], with 10.5% and 11.2% improvements in the gold metric BACC, respectively. Meanwhile, our model also outperforms the transformer-based methods, such as ViT-B/16 and Swin-T. Compared with the conventional Swin-T, our top-down method incorporating LSM improves Acc by 1.4% after fusing global and lesion information.

Since metadata are also available for the cases in the PAD-UFES-20 dataset, we compare our model involving both the clinical image branch and metadata branch with other multi-modal models, the results of which are presented in the lower part of Table III denoted with CM in Modality column. Adding metadata significantly improves the accuracy of the classification for all methods. Comparing our method with only clinical image branches, an improvement of 5.0%, 3.4%, and 7.2% in BACC, AUC, and Acc is obtained, which justifies the effectiveness of metadata and CMF. Similar to our clinical image branch, the proposed method still achieves leading BACC over other methods with multi-modal inputs, which proves the superiority of the overall design.

B. Comparison with SOTA methods on ISIC datasets

To assess the efficacy of the dermoscopy branch, we conduct two distinct tasks on ISIC datasets: a disease diagnosis task similar to the CI-Net [30], and an attribute recognition task as the GIIN [33]. As shown in Table IV our bottom-up method achieves more competitive performance on both ISIC 2018 and 2019 datasets by 5-fold cross-validation, where the results of other methods are obtained from [30]. Our two-level structure extracts the fine-grained textural feature from the original resolution, which implicitly realizes the zooming function rather than explicitly implementing the zooming and comparing counterparts as in [30]. Moreover, the MSTA in our dermoscopy branch has the memory ability to store learned texture templates serving the role of the comparing step similar

TABLE IV: Comparison of our dermoscopy image branch with other methods on ISIC 2018 and ISIC 2019 datasets. The results of other methods are sourced from [30]. *BMCA is the gold metric (%).

Dataset	Method	AP	Acc	AUC	*BMCA
ISIC 2018	EfficientNet [32]	86.3	92.3	97.1	83.7
	Taxonomies [49]	85.0	93.4	97.3	83.7
	L-CNN [50]	87.4	94.4	98.2	84.4
	MB-DCNN [28]	85.7	93.9	94.5	83.2
	CI-Net [30]	88.3	**95.7**	97.0	85.5
	Ours-D	**93.9**	94.1	**99.2**	**87.0**
ISIC 2019	EfficientNet [32]	53.1	83.5	87.1	63.9
	Taxonomies [49]	51.2	84.0	86.7	63.2
	L-CNN [50]	53.4	81.3	86.3	63.2
	MB-DCNN [28]	52.1	81.4	86.9	64.3
	CI-Net [30]	54.1	**84.9**	88.2	65.7
	Ours-D	**78.2**	81.7	**96.3**	**72.0**

TABLE V: Performance (AUC) of the different models on attribute recognition on the ISIC 2017 and ISIC 2018 datasets (%). Results of ResNet50 and GIIN are obtained from [33].

Dataset	Method	Attribute					Avg.
		MLCs	NPN	PN	STR	GLO	
ISIC 2017	ResNet50	62.6	78.8	86.2	**97.1**	-	81.2
	GIIN	68.3	84.1	87.4	93.9	-	83.4
	Ours-D	**74.7**	**86.8**	**89.6**	95.6	**-**	**86.6**
ISIC 2018	ResNet50	58.7	84.1	89.0	87.6	67.8	77.5
	GIIN	68.8	87.3	86.4	88.7	77.1	81.6
	Ours-D	**74.3**	**87.8**	**91.2**	**91.8**	**82.7**	**85.6**

in [30]. The comparison results demonstrate our approach can generate better texture features, consequently yielding an enhancement in performance. To be specific, compared to CI-Net, our model has gained an improvement of BMCA by 1.5% and 6.3% on ISIC 2018 and ISIC 2019, respectively.

It can be attributed to the fact that many key attributes of skin tumor diagnosis, such as stripes, globules, and pigment networks are formed by subtle textural patterns. A better representation of these textures can lead to enhanced attribute recognition to effectively mitigate the long-tailed issue [48]. Thus, we conduct experiments on attribution recognition tasks by comparing them with the prior state-of-the-art method GIIN. In terms of attribute recognition, Table V demonstrates that our method has a significant advantage, with an average AUC improvement of 3.2% and 4% on ISIC 2017 and ISIC 2018, respectively.

C. Comparison with SOTA methods on Derm7pt dataset

We benchmark RemixFormer++ with state-of-the-art methods on high-quality multi-modal dataset Derm7pt. Table VI lists the accuracy of eight classification tasks, indicating that our RemixFormer++ performs well for all checklist labels, particularly in DIAG classification. As can be observed, our CDM model gives the best accuracy on DIAG classification with 4%, 8.3%, and 13.9% increments compared to FM4Net, Incep-co, and EmbNet, respectively. Simultaneously, our CDM model also outperforms all the comparison methods on average accuracy (Avg.). Since Triple-Net, HcCNN, and AMFAM only use clinical and dermoscopy images, for a matter of fairness, we also report the result of Remixformer++ with clinical and dermoscopy images only, in the row denoted as CD in the Modality column. As shown, our CD model also presents significant improvements over Triple-Net, HcCNN, and AMFAM

TABLE VI: Comparison of our model with state-of-the-art methods across different modalities for accuracy on the Derm7pt dataset, where C, D, and M represent clinical images, dermoscopy images, and patient metadata respectively. Except for our model and FM4Net [9], other methods did not report results for all different modalities. Avg. represents the average accuracy of disease diagnosis (DIAG) and 7-point checklist labels (%).

Method	Modality	7-Point Checklist							DIAG	Avg.
		BWV	DaG	PIG	PN	RS	STR	VS		
TripleNet [6]	CD	87.9	61.3	67.3	63.3	76.0	74.4	83.0	68.6	72.7
HcCNN [13]	CD	87.1	65.6	68.6	70.6	80.8	71.6	**84.8**	69.9	74.9
EmbNet [45]	CDM	84.3	57.5	64.3	65.1	78.0	73.4	82.5	68.6	71.7
Incep-co [11]	CDM	87.1	60.0	66.1	70.9	77.2	74.2	79.7	74.2	73.7
AMFAM [46]	C	83.0	53.7	64.1	55.2	72.4	66.3	80.0	64.8	67.4
	D	87.1	60.0	66.6	66.1	78.5	71.1	80.5	69.4	72.4
	CD	**88.1**	63.8	70.9	70.6	80.8	74.7	83.3	75.4	76.0
FM4Net [9]	C	83.7	53.3	59.3	57.5	74.5	66.3	80.8	67.0	67.8
	D	87.2	60.0	68.3	69.0	80.1	73.7	81.4	74.7	74.3
	CD	87.9	60.2	**71.5**	67.6	81.9	73.8	82.1	75.6	75.1
	CM	84.8	62.1	60.2	64.7	73.9	71.9	79.3	70.5	70.9
	DM	87.6	66.2	69.7	69.3	80.2	75.0	80.6	76.6	75.6
	CDM	**88.1**	66.1	70.1	71.1	81.5	**78.0**	81.8	78.5	77.0
Ours	C	82.3	52.9	59.5	59.5	72.2	65.3	79.5	70.6	67.7
	D	85.1	61.8	67.6	**74.7**	79.8	72.4	83.3	75.4	75.0
	CD	87.1	62.8	70.4	**74.7**	**83.8**	72.9	82.0	77.7	76.4
	CM	84.1	57.7	53.9	65.1	71.4	69.6	77.7	73.9	69.2
	DM	86.1	63.3	66.1	73.4	80.8	72.9	84.1	80.3	75.9
	CDM	86.3	**66.6**	70.6	**74.7**	82.0	72.2	83.3	**82.5**	**77.3**

TABLE VII: Ablation study for RemixFormer++ in terms of accuracy on Derm7pt dataset (%).

Method	Backbone	7-Point Checklist							DIAG	Avg.
		BWV	DaG	PIG	PN	RS	STR	VS		
Ours-C	Swin-T	82.8	55.7	58.5	54.4	72.6	65.1	81.0	66.8	67.1
	Swin-T+LSM	81.5	53.9	57.7	57.5	71.9	68.4	80.0	69.1	67.5
	Swin-T+LSM+CSF	82.3	52.9	59.5	59.5	72.2	65.3	79.5	70.6	67.7
Ours-D	ViT-S/16	85.6	56.5	65.1	57.5	76.2	62.0	79.0	65.6	68.4
	ViT-S/16+HRWE	85.8	63.0	67.9	67.6	81.3	72.6	84.1	75.2	74.7
	ViT-S/16+HRWE+MSTA	85.1	61.8	67.6	74.7	79.8	72.4	83.3	75.4	75.0

by 9.1%, 7.8%, and 2.3%, respectively, which convince the effectiveness of our clinical image branch and dermoscopy image branch. For the sake of completeness, we evaluate our method in different modality combinations and compare them with the existing methods whenever those combinations are available. The results in Table VI illustrate that the clinical image, dermoscopy image, and metadata are complementary features for skin lesion diagnosis. By effectively combining the multi-modality information, it can significantly improve diagnostic accuracy.

In what follows, we use an ablation study to analyze the contribution of LSM and CSF modules in the clinical image branch, as well as the contribution of HRWE and MSTA modules in dermoscopy image branch. As shown in Table VII, the incorporation of the LSM and CSF modules leads to a progressive improvement in both the DIAG and average (Avg.) accuracy for Ours-C model, evidencing their contribution in the final diagnosis. A comparable enhancement is also observed in Ours-D model. It is worth emphasizing that the HRWE module can significantly improve diagnostic accuracy with around 10% improvement in DIAG. Due to the improvements in clinical and dermoscopy branches, the proposed methods achieve a new record of 82.5% accuracy in CDM, surpassing our previous leading work [15]; see Table VIII.

Table VIII provides quantitative information on the recognition performance of five diseases and seven attributes related to Mel. These results demonstrate that RemixFormer++ performs well on various metrics (Sens, Spec, AUC, and F1) in both CD and CDM. Compared to [15], the model's average F1 in DIAG improves by 5.3% in CDM. Furthermore, RemixFormer++ demonstrates competitive performance in attribute recognition, with an average F1 improvement of 2.4% compared to [45].

D. GPU Memory Consumption

We assessed GPU memory consumption for our clinical branch (Ours-C) and dermoscopy branch (Ours-D) with ViT-S/16 and Swin-T, using different batch sizes and image resolutions. The comparisons are summarized in Table IX. In actual experiments, we primarily used 2048^2 resolution for dermoscopy images and 896^2 resolution for clinical images with 224^2 lesion patch in LSM. For a broader comparison, we selected two standard resolutions, 512^2 and 2048^2, to analyze GPU consumption.

We use the command "nvidia-smi" and the function "torch.cuda.max_memory_allocated" to track the peak GPU memory usage and maximal GPU memory allocated by all tensors, referred to as nvidia-smi and max memory, respectively. Table IX illustrates that with a 512^2 image resolution and batch size of 1, Ours-D only requires 1235MiB GPU memory for training and 194MiB for tensors, which is notably lower than ViT-S and Swin-T. Similarly, Ours-C consumes 1265MiB of GPU memory, less than both ViT-S and Swin-T. Doubling the batch size, our models maintain their efficiency as feature map

TABLE VIII: Comparison of RemixFormer++ with other methods on the DIAG and 7-point checklist classification task on Derm7pt (%). Avg1 represents the average metrics corresponding to 5 categories of DIAG, and Avg2 represents the average metrics of 8 categories, which include Mel and 7 attributes that are positively associated with a diagnosis of melanoma. Remix is only used for DIAG classification in the previous paper [15].

Met.	Mod.	Method	DIAG					Avg1	PN	STR	PIG	RS	DaG	BWV	VS	Avg2
			BCC	Nev	Mel	Misc	SK		ATP	IR	IR	PRS	IR	PRS	IR	
Sens	CD	TripleNet [6]	-	-	46.5	-	-	-	33.3	39.4	61.3	97.9	67.2	90.0	30.0	58.2
		HcCNN [13]	-	-	58.4	-	-	-	40.9	35.1	55.7	95.2	80.2	92.2	20.0	**59.7**
		AMFAM [46]	40.0	84.1	65.8	68.0	40.0	59.6	58.5	57.3	67.9	72.1	66.7	75.0	0.0	57.9
		GIIN [33]	-	-	59.0	-	-	-	77.5	67.0	39.2	21.9	70.1	69.9	3.6	51.0
		Ours-CD	81.3	87.7	66.3	65.0	47.4	**69.5**	58.1	59.6	61.3	52.8	70.1	72.0	30.0	58.8
	CDM	Incep-co [11]	62.5	88.6	61.4	47.5	42.1	60.4	48.4	51.1	59.7	66.0	62.1	77.3	13.3	54.9
		EmbNet [45]	-	-	40.6	-	-	-	33.3	51.1	60.5	96.2	64.4	96.3	23.3	58.2
		FM4Net [9]	43.8	95.0	71.3	52.5	10.5	54.6	49.5	55.3	54.8	44.3	76.3	64.0	0.3	52.0
		Remix [15]	68.8	92.7	73.3	65.0	36.8	67.3	-	-	-	-	-	-	-	-
		Ours-CDM	93.8	90.0	72.3	75.0	57.9	**77.8**	61.3	67.0	63.7	47.2	76.8	72.0	33.3	**61.7**
Spec	CD	TripleNet [6]	-	-	90.1	-	-	-	93.0	92.4	76.8	23.6	71.6	60.0	98.6	75.8
		HcCNN [13]	-	-	88.1	-	-	-	92.4	90.0	86.3	41.5	71.6	65.3	98.4	79.2
		AMFAM [46]	97.4	85.8	91.4	93.8	95.6	**92.8**	85.6	85.9	83.0	82.6	82.4	90.3	92.4	86.7
		GIIN [33]	-	-	89.5	-	-	-	79.0	80.3	95.8	96.8	78.8	91.0	100.0	**88.9**
		Ours-CD	96.6	79.5	90.1	98.0	99.2	92.7	89.1	83.1	84.1	95.2	78.0	90.6	96.4	88.3
	CDM	Incep-co [11]	97.9	71.6	88.8	97.5	99.5	91.1	90.7	85.7	80.1	81.3	78.9	89.4	97.5	86.6
		EmbNet [45]	-	-	93.5	-	-	-	90.7	88.7	82.7	20.8	78.4	52.0	96.7	75.4
		FM4Net [9]	97.9	73.3	91.5	98.9	99.7	92.3	90.7	89.7	89.7	95.2	76.6	94.4	99.7	**90.9**
		Remix [15]	97.4	78.4	92.9	98.9	99.7	93.4	-	-	-	-	-	-	-	-
		Ours-CDM	96.8	85.8	93.9	96.9	99.2	**94.5**	89.7	79.7	84.5	94.8	75.2	89.7	97.3	88.1
AUC	CD	TripleNet [6]	-	-	81.2	-	-	-	73.8	76.3	77.6	76.8	76.0	85.1	79.9	78.4
		HcCNN [13]	-	-	85.6	-	-	-	78.3	77.6	81.3	81.9	82.6	89.8	82.7	82.5
		AMFAM [46]	94.1	89.7	**89.1**	90.6	81.7	89.0	82.0	80.7	83.4	86.7	81.9	91.1	80.9	**84.5**
		GIIN [33]	92.8	86.8	87.6	88.8	79.8	87.2	87.5	81.2	83.6	79.0	83.1	90.8	75.4	83.5
		Ours-CD	97.7	89.5	86.9	95.2	91.2	**92.1**	85.9	79.0	83.6	84.6	81.5	91.4	83.6	**84.5**
	CDM	Incep-co [11]	92.9	89.7	86.3	88.3	91.0	89.6	79.9	78.9	79.0	82.9	79.9	89.2	76.1	81.5
		EmbNet [45]	-	-	82.5	-	-	-	74.5	77.7	77.9	71.3	78.5	84.8	76.9	78.0
		FM4Net [9]	95.4	94.6	92.6	95.0	91.5	93.8	85.7	84.6	84.9	83.0	84.5	92.5	81.3	86.1
		Remix [15]	96.1	94.2	92.0	95.7	93.8	**94.4**	-	-	-	-	-	-	-	-
		Ours-CDM	96.3	92.6	91.3	96.2	94.1	94.1	87.7	81.6	82.5	86.1	84.2	91.1	86.2	**86.3**
F1	CD	TripleNet [6]	-	-	53.1	-	-	-	42.7	48.1	57.8	86.6	66.4	90.3	40.9	60.7
		HcCNN [13]	-	-	60.5	-	-	-	49.4	42.0	60.0	87.9	74.5	92.0	28.6	**61.9**
		AMFAM [46]	38.7	86.7	70.6	52.3	16.0	52.9	54.8	55.8	64.4	56.3	73.8	64.1	0.0	55.0
		GIIN [33]	-	-	62.1	-	-	-	59.6	57.5	53.1	33.7	72.4	68.6	6.9	51.8
		Ours-CD	61.9	85.9	68.0	71.2	58.1	**69.0**	60.0	55.7	62.6	63.6	71.1	67.9	34.6	60.4
	CDM	Incep-co [11]	58.8	83.8	63.3	55.9	55.2	63.4	54.2	51.9	58.7	60.9	66.0	69.4	18.6	55.4
		EmbNet [45]	-	-	50.9	-	-	-	40.8	54.6	61.0	85.4	67.4	92.8	28.5	60.2
		FM4Net [9]	45.2	87.8	72.7	64.6	18.1	57.7	55.1	58.8	61.8	56.2	74.4	68.1	0.6	56.0
		Remix [15]	59.5	88.3	75.5	74.3	51.9	69.9	-	-	-	-	-	-	-	-
		Ours-CDM	69.8	89.3	76.0	74.1	66.7	**75.2**	63.0	57.8	64.5	58.5	74.1	66.7	40.0	**62.6**

memory consumption increases. At a resolution of 2048^2, the advantage of our models is even more pronounced, using less memory in total than either Swin-T or ViT-S individually. To sum up, the remarkable memory efficiency of our method is particularly useful when handling high-resolution images.

E. Ablation Study

We evaluate the impact of the proposed LSM and HRWE components by ablation study. All experiments are performed with 5-fold cross-validation on the large-scale X-SkinTumor-12 for consistency, and we also perform a paired t-test for comparing F1 between the referenced baseline (Ref.) with different methods.

1) Effectiveness of Lesion Selection Module: We conduct a study using clinical images to explore how integrating the LSM module into various backbones affects the diagnosis of skin tumors. Table X shows that utilizing LSM to combine global and local features can be advantageous for Swin-T, ViT-S/16 and ViT-B/16 models. Specifically, for Swin-T, LSM significantly (p=0.0172) enhances F1 by 1.9%. Additionally, we find that CSF consistently improves classification results across different backbones. Our Swin-T-based model also

TABLE IX: Comparison of GPU memory consumption w.r.t. different Input image sizes and Batch sizes.

Method	BatchSize	InputSize	GPU Memory Usage	
			nvidia-smi	max memory
ViT-S	1	512×512	1347MiB	268MiB
Swin-T	1	512×512	1305MiB	204MiB
Ours-C	1	512×512	1265MiB	241MiB
Ours-D	1	512×512	1235MiB	194MiB
ViT-S	2	512×512	1411MiB	351MiB
Swin-T	2	512×512	1473MiB	299MiB
Ours-C	2	512×512	1343MiB	291MiB
Ours-D	2	512×512	1291MiB	224MiB
ViT-S	1	2048×2048	18187MiB	16906MiB
Swin-T	1	2048×2048	3469MiB	1675MiB
Ours-C	1	2048×2048	2299MiB	1000MiB
Ours-D	1	2048×2048	1809MiB	646MiB
ViT-S	2	2048×2048	36358MiB	33865MiB
Swin-T	2	2048×2048	5691MiB	3170MiB
Ours-C	2	2048×2048	3403MiB	1779MiB
Ours-D	2	2048×2048	2411MiB	1128MiB

TABLE X: Ablation study for our clinical image branch on X-SkinTumor-12 dataset (%).

Method	Backbone	#param.	LSM	CSF	Sens	Spec	Prec	AUC	Acc	F1	p-value
Ours-C	ViT-S/16	48.3M			62.2±1.2	97.7±0.0	69.7±1.5	95.0±0.5	79.9±0.5	64.6±0.8	Ref.
	ViT-S/16	48.9M	✓		62.5±1.4	97.7±0.0	71.3±2.3	94.8±0.6	80.3±0.3	65.0±1.2	0.0684
	ViT-S/16	53.6M	✓	✓	63.2±0.5	97.8±0.0	70.6±2.8	94.8±0.6	80.5±0.3	65.5±0.8	0.0427
	ViT-B/16	86.1M			67.6±1.6	98.1±0.1	73.8±2.7	**97.2±0.4**	82.6±0.7	69.6±1.4	0.0001
	ViT-B/16	86.7M	✓		68.1±0.4	98.1±0.1	72.9±3.3	96.9±0.3	82.9±0.5	69.6±1.2	4.8e-05
	ViT-B/16	91.4M	✓	✓	67.4±1.0	98.1±0.1	76.0±2.7	96.5±0.6	83.1±0.5	70.2±1.1	1.5e-05
	Swin-T	27.9M			64.8±0.8	98.0±0.1	72.1±3.0	96.7±0.4	81.9±0.6	67.4±1.4	Ref.
	Swin-T	43.6M	✓		67.3±1.7	98.1±0.1	72.9±2.4	96.8±0.6	83.1±0.8	69.3±1.7	0.0172
	Swin-T	48.3M	✓	✓	**68.1±1.3**	**98.2±0.1**	**74.3±1.0**	**97.2±0.5**	**83.6±0.7**	**70.5±0.8**	0.0012

TABLE XI: Ablation study of the selected patch number K_p in the LSM module (%).

Method	Backbone	LSM	Sens	Spec	Prec	AUC	Acc	F1
Ours-C	Swin-T	K_p=2	65.8±1.8	98.0±0.1	73.0±2.0	96.7±0.4	82.3±0.8	68.1±1.5
	Swin-T	K_p=4	**68.1±1.3**	**98.2±0.1**	74.3±1.0	**97.2±0.5**	**83.6±0.7**	**70.5±0.8**
	Swin-T	K_p=6	68.0±1.5	98.1±0.1	**75.0±2.6**	97.0±0.5	83.3±1.0	70.2±1.5
	Swin-T	K_p=8	66.8±1.5	98.1±0.1	73.8±2.8	97.0±0.3	83.2±0.8	69.1±1.8

TABLE XII: Ablation study for different backbones in our dermoscopy image branch (%).

Method	Backbone	#param.	Sens	Spec	Prec	AUC	Acc	F1
Ours-D	Swin-T	27.5M	65.3±2.8	98.6±0.0	72.1±2.8	98.4±0.4	89.3±0.4	67.3±1.5
	Swin-B	89.6M	**74.0±0.4**	**98.8±0.1**	73.3±2.8	98.1±0.2	89.4±1.1	70.2±2.5
	Swin-T+HRWE	49.1M	64.0±2.1	98.5±0.1	71.5±2.5	98.3±0.3	88.4±0.7	65.5±1.9
	ViT-S/16	48.3M	63.3±3.6	98.5±0.1	68.9±2.6	98.0±0.3	88.3±0.4	64.8±2.4
	ViT-B/16	86.1M	68.2±1.9	98.6±0.1	71.1±2.8	98.4±0.1	89.3±0.8	68.3±2.8
	ViT-S/16+HRWE	43.0M	71.5±1.7	**98.8±0.1**	**76.1±4.2**	**98.7±0.4**	**90.7±0.7**	**72.5±3.0**

TABLE XIII: Ablation study for dermoscopy image branch on X-SkinTumor-12 (%).

Method	Backbone	ImgSize	Weight		HRWE	MSTA	Sens	Spec	Prec	AUC	Acc	F1	p-value
			IN	SSL									
Ours-D	ViT-S/16	384^2	✓				63.3±3.6	98.5±0.1	68.9±2.6	98.0±0.3	88.3±0.4	64.8±2.4	Ref.
	ViT-S/16	512^2	✓				65.5±3.2	98.6±0.0	71.1±3.3	98.0±0.4	89.1±0.5	67.2±1.9	0.0541
	ViT-S/16	512^2		✓			67.4±0.3	98.6±0.1	73.7±4.7	98.6±0.1	89.4±0.5	69.4±3.4	0.0437
	ViT-S/16	2048^2	✓		✓		64.0±1.8	98.5±0.1	67.2±3.1	97.3±0.7	88.3±0.6	64.4±1.9	0.9113
	ViT-S/16	2048^2		✓	✓		71.5±1.7	98.8±0.1	76.1±4.2	98.7±0.4	90.7±0.7	72.5±3.0	5.2e-05
	ViT-S/16	2048^2		✓	✓	✓	**74.1±2.6**	98.9±0.1	**77.3±3.4**	**98.9±0.3**	**91.2±0.7**	**74.2±2.4**	0.0003

TABLE XIV: Modality-wise ablation study for RemixFormer++ in terms of accuracy on X-SkinTumor-12 (%).

Method	#param.	Sens	Spec	Prec	AUC	Acc	F1
Ours-C	48.3M	58.3±0.4	97.7±0.1	68.1±0.9	96.8±0.4	80.7±0.8	61.6±0.7
Ours-D	43.0M	70.6±1.0	98.6±0.1	79.4±1.0	98.0±0.3	87.8±0.5	73.7±0.8
Ours-CM	59.4M	72.0±0.3	98.9±0.0	78.7±0.3	99.1±0.0	89.9±0.2	74.4±0.4
Ours-CD	76.6M	73.5±1.4	98.8±0.0	78.3±0.8	98.7±0.1	89.9±0.3	75.0±0.8
Ours-DM	47.8M	74.2±0.9	98.9±0.0	82.0±1.3	98.7±0.0	90.5±0.2	76.9±0.8
Ours-CDM	76.7M	81.5±1.4	99.2±0.0	84.6±0.6	99.4±0.0	92.6±0.2	82.4±0.8

outperforms similar size ViT-S/16 and much larger ViT-B/16 models. Moreover, an ablation study is performed on the value of $K_p = 2, 4, 6, 8$ to determine the optimal value of K_p for lesion selection. As indicated in Table XI, it can be observed that $K_p = 4$ gives the best result as it produces enough coverage for the lesion area and avoids bringing too much context information into the local feature encoder. Thus, we choose $K_p = 4$ in all the experiments.

2) Effectiveness of High-Resolution Window-Level Encoder: In the dermoscopy branch, we compare the single-level method with various backbones (Table XII) to our two-level architecture with the novel HRWE. Larger backbones (Swin-B, ViT-B) consistently outperform smaller ones in single-level models, with Swins showing better performance than ViTs. However, our two-level approach (Swin-T+HRWE, ViT-S/16+HRWE) reveals different phenomena. Swins excel at capturing higher-level features, but Swin-T in the first level underperforms due to potential information loss. Conversely, our two-level ViT-S/16+HRWE achieves the best results with 90.7% accuracy and 72.5% F1 score. Its compact size also makes it suited for memory-intensive multi-modal settings, being the choice of our dermoscopy branch's backbone. We also conduct experiments to assess the impact of different input sizes and the HRWE. Table XIII illustrates that using a higher resolution (512^2) can increase F1 by 2.4% compared to 384^2 in ViT-S/16. Furthermore, when employing a two-level hierarchical architecture for higher resolution images (2048^2), HRWE leads to a significant (p=5.2e-5) performance improvement on F1 compared to the single-level ViT-S/16. We demonstrate the importance of using self-supervised pre-trained weights (SSL) through comparative experiments in both single-level and two-level methods. The T_{ϕ_1} model in HRWE using SSL weights

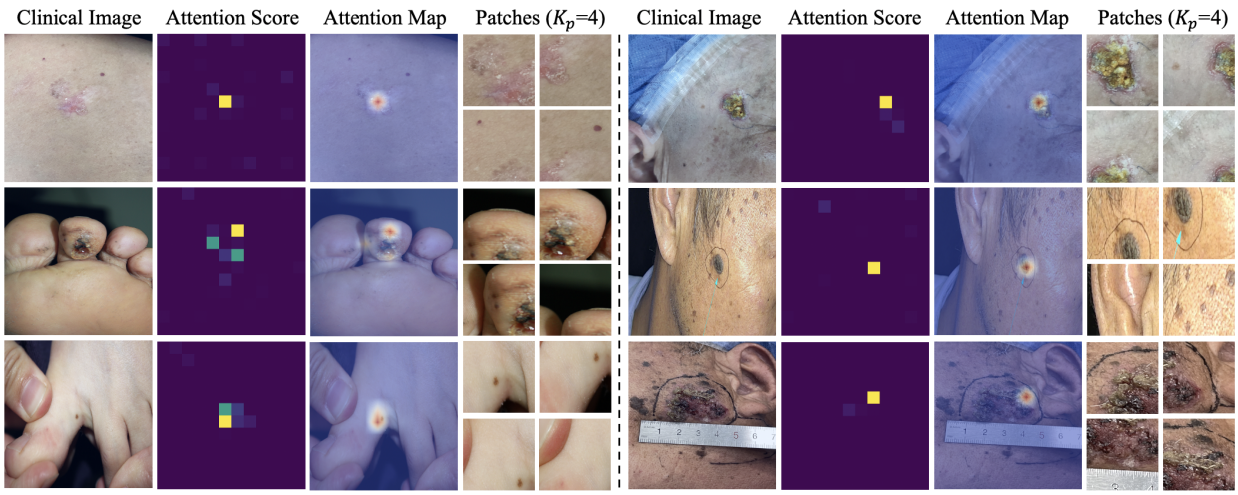


Fig. 7: Visualization of patch (localization) maps for the clinical branch on X-SkinTumor-12 dataset, where the attention score and attention map are shown in accord with each other.

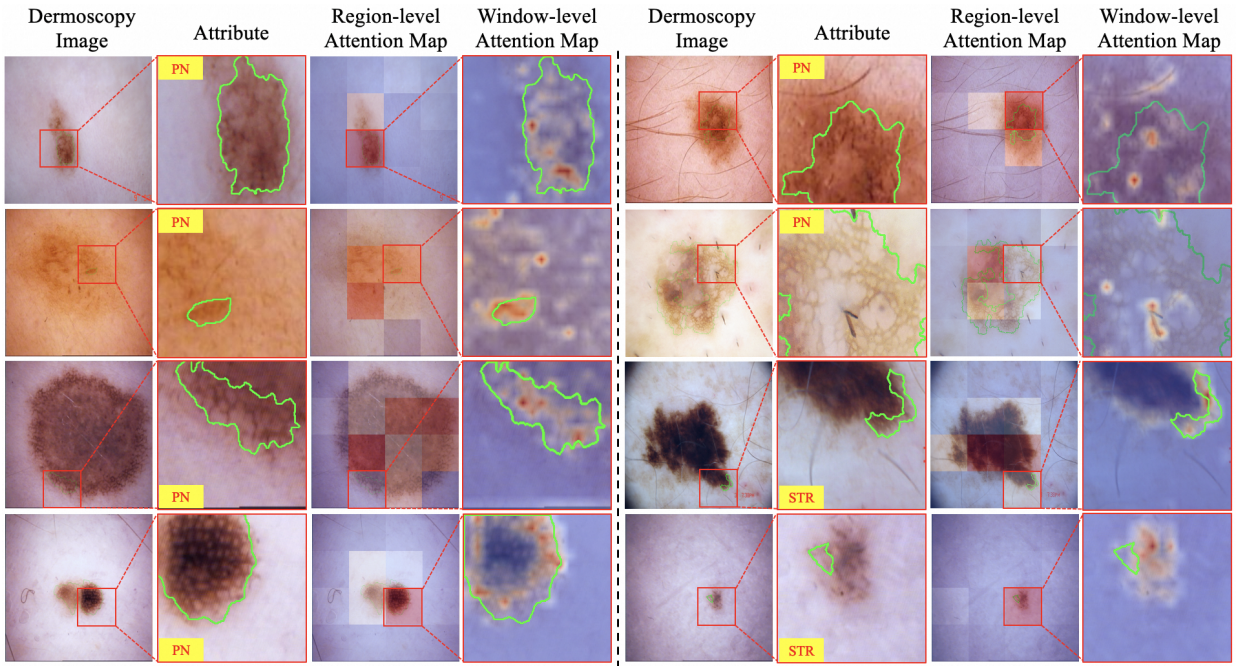


Fig. 8: Visualization of region-level and window-level attention maps for the dermoscopy branch, where the good cases and bad cases are listed on the left side and right side, respectively. Note that dermoscopic attributes are highlighted with green outlines.

outperforms the one using ImageNet weights (IN), with 2.4% and 8.1% improvements in Acc and F1. Additionally, MSTA adding texture embedding to window embedding provides an extra 1.7% gain in F1.

3) Attention Map Analysis: Our approach also provides attention outputs from the LSM and HRWE, facilitating an automatic assessment of patch importance and a visual verification for the confidence of the final diagnosis. We visualize some typical attention scores and attention maps from the clinical branch in Fig. 7 and dermoscopy branch in Fig. 8. Specifically, Fig. 7 illustrates that the lesion areas have the highest attention values, providing compelling evidence that our global feature encoder effectively focuses on the relevant regions. Thus, our LSM has demonstrated its capability to extract valuable local features for feature fusion. Likewise, HRWE also aims to better extract texture information and learn high-resolution dermoscopy image representations. Fig. 8 indicates that the HRWE is effective at capturing important visual features for skin tumor diagnosis. The attention maps on the left examples can accurately locate the target regions for typical lesions, highly consistent with the attributes. However, challenging cases on the right examples failed to highlight the lesion region due to interference factors such as hair and blood vessels. Thus, we still need to further improve our model to detect such atypical lesion areas.

4) Effectiveness of Multi-modality Data for Skin Tumor Diagnosis: In our study, we highlight the significance of using multi-modal data in skin tumor diagnosis. We select 1500

Disease	Sex		Color					Location							Evolution				
	Male	Female	Yellow	Brown	Black	Red	Other	Face	Neck	Trunk/ Limbs	Genitals	Head	Breast	Other	Trauma	Born	Slowly growing	Rapidly growing	None
BCC	0.443	0.408	0.050	0.149	0.591	0.368	0.001	0.847	0.013	0.071	0.016	0.047	0.003	0.008	0.129	0.006	0.458	0.069	0.033
Mel	0.584	0.241	0.038	0.195	0.753	0.173	0.011	0.036	0.000	0.907	0.008	0.000	0.000	0.047	0.099	0.033	0.625	0.104	0.033
SCC	0.596	0.383	0.057	0.267	0.394	0.623	0.000	0.509	0.008	0.288	0.070	0.092	0.000	0.032	0.291	0.003	0.666	0.005	0.013
Bowen	0.232	0.293	0.067	0.628	0.186	0.537	0.006	0.143	0.009	0.271	0.168	0.012	0.000	0.387	0.070	0.021	0.320	0.012	0.018
Paget	0.779	0.216	0.101	0.085	0.015	0.884	0.000	0.005	0.000	0.040	0.864	0.000	0.065	0.020	0.141	0.000	0.714	0.000	0.000
AK	0.335	0.519	0.171	0.557	0.051	0.367	0.013	0.905	0.003	0.016	0.000	0.060	0.000	0.013	0.060	0.000	0.456	0.000	0.190
Kel	0.378	0.616	0.200	0.103	0.016	0.676	0.000	0.173	0.016	0.568	0.086	0.032	0.000	0.119	0.443	0.086	0.162	0.005	0.335
DF	0.379	0.614	0.150	0.536	0.143	0.179	0.007	0.050	0.029	0.800	0.007	0.014	0.000	0.114	0.143	0.200	0.179	0.007	0.500
SN	0.516	0.442	0.335	0.284	0.088	0.312	0.009	0.298	0.042	0.042	0.009	0.609	0.000	0.009	0.065	0.623	0.228	0.000	0.088
SK	0.346	0.649	0.081	0.473	0.446	0.053	0.000	0.418	0.089	0.364	0.033	0.092	0.005	0.059	0.223	0.057	0.450	0.003	0.249
Nev	0.364	0.581	0.073	0.588	0.320	0.056	0.003	0.388	0.083	0.426	0.021	0.053	0.015	0.035	0.002	0.224	0.075	0.003	0.022
Hem	0.445	0.536	0.018	0.045	0.182	0.709	0.036	0.282	0.045	0.427	0.036	0.091	0.027	0.091	0.018	0.627	0.255	0.045	0.173

Disease	Sign					Age of onset					Duration			Medical history			Sun exposure time		
	Pain	Itching	Bleeding	No hair	None	<20	20-30	31-40	41-50	>50	<5 years	5-10 years	>10 years	Yes	No	Unknown	<2 h/day	2-5 h/day	>5 h/day
BCC	0.081	0.130	0.246	0.006	0.329	0.034	0.017	0.120	0.244	0.265	0.384	0.171	0.129	0.120	0.409	0.133	0.436	0.116	0.058
Mel	0.282	0.197	0.238	0.008	0.304	0.134	0.027	0.211	0.148	0.279	0.151	0.274	0.370	0.099	0.490	0.205	0.562	0.142	0.090
SCC	0.022	0.019	0.350	0.000	0.596	0.019	0.008	0.070	0.547	0.332	0.836	0.100	0.040	0.049	0.663	0.264	0.749	0.224	0.003
Bowen	0.000	0.052	0.235	0.000	0.159	0.000	0.009	0.110	0.204	0.119	0.369	0.067	0.006	0.064	0.317	0.061	0.320	0.122	0.000
Paget	0.065	0.437	0.553	0.000	0.271	0.000	0.000	0.101	0.256	0.497	0.638	0.191	0.025	0.216	0.427	0.211	0.618	0.236	0.000
AK	0.070	0.168	0.253	0.006	0.225	0.016	0.006	0.032	0.130	0.516	0.275	0.405	0.019	0.076	0.275	0.348	0.022	0.158	0.519
Kel	0.232	0.319	0.076	0.032	0.405	0.378	0.416	0.146	0.038	0.011	0.649	0.303	0.038	0.216	0.486	0.286	0.724	0.259	0.005
DF	0.129	0.221	0.029	0.021	0.593	0.043	0.586	0.271	0.057	0.036	0.621	0.300	0.071	0.157	0.543	0.293	0.864	0.129	0.000
SN	0.023	0.121	0.014	0.140	0.628	0.791	0.042	0.028	0.009	0.033	0.321	0.209	0.372	0.033	0.688	0.181	0.726	0.177	0.000
SK	0.096	0.310	0.003	0.000	0.572	0.019	0.080	0.563	0.132	0.186	0.577	0.311	0.092	0.142	0.627	0.211	0.912	0.067	0.001
Nev	0.002	0.008	0.000	0.001	0.265	0.255	0.010	0.004	0.001	0.004	0.116	0.068	0.091	0.002	0.256	0.017	0.242	0.030	0.003
Hem	0.073	0.118	0.018	0.000	0.773	0.673	0.164	0.064	0.018	0.064	0.400	0.218	0.364	0.045	0.745	0.191	0.836	0.145	0.000

Fig. 9: Statistical analysis of varied attributes present in the metadata of the X-SkinTumor-12 dataset. The intensity of the color in the grid represents the proportion of samples, with darker shades indicating a higher percentage. The chart reveals no significant gender-based differences in the distribution of samples across various diseases, except for Paget's disease. Compared with benign tumors, most malignant tumors such as BCC, Mel and SCC tend to grow slower and are often accompanied by symptoms such as pain, itching, and bleeding. Additionally, prolonged sun exposure is strongly associated with AK.

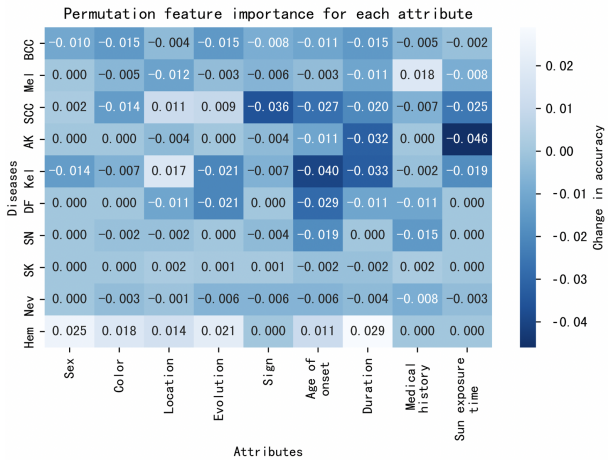


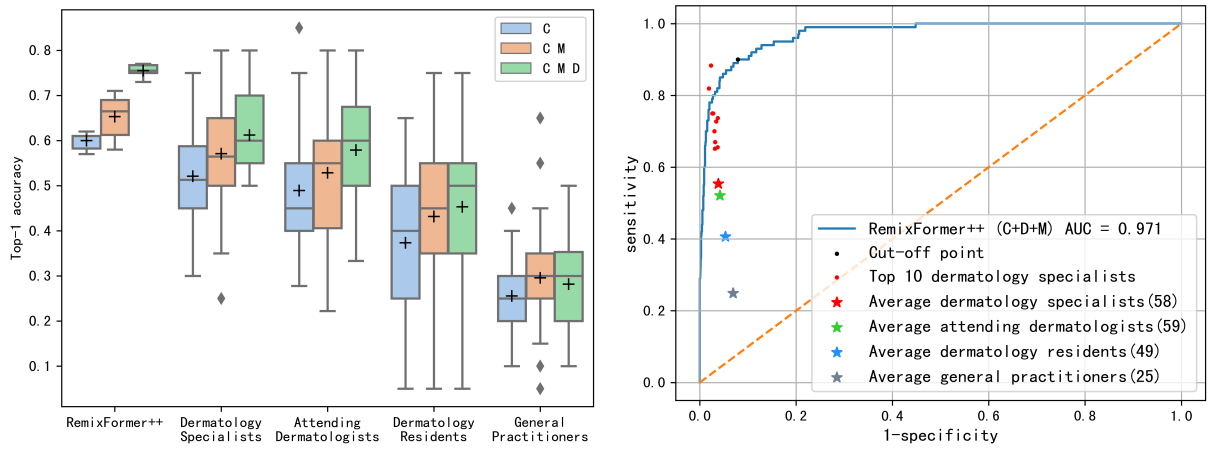
Fig. 10: The heatmap visualizes the permutation feature importance for each attribute in predicting diseases. Each cell quantifies the change in model accuracy when the corresponding attribute is perturbed, serving as an indicator of the attribute's importance. Darker colors indicate higher importance, showing which ones have a stronger impact on diagnosis.

paired cases from X-SkinTumor-12 as the test set, and conduct modality-wise ablation experiments, as shown in Table XIV. Our findings indicate that incorporating two modalities (CD/CM/DM) leads to a significant improvement in performance compared to using a single modality (i.e., C or D). By combining all three modalities (i.e., CDM), we can achieve an overall accuracy of 92.6% on the X-SkinTumor-12 dataset.

5) Importance of Clinical Metadata for Skin Tumor Diagnosis: We conduct experiments using 1500 paired cases from the X-SkinTumor-12 dataset as test samples to explore the impact of the 9 metadata attributes on the multi-modal diagnostic model. Specifically, we employ the permutation feature importance method by randomly altering the value of a certain attribute and calculating the deviation in model accuracy caused by this change. The deviation reflects the attribute's significance. Larger deviations indicate a greater impact on the model. Fig. 10 illustrates the findings, revealing that age of onset, sign, duration, and sun exposure time have a substantial impact on accuracy, which aligns with the actual weight distribution of the real metadata (Fig. 9). For instance, the perturbing attribute "Sun exposure time" shows a relatively significant impact on the model's accuracy for predicting AK, resulting in an accuracy decrease of 0.046. Similarly, the "Sign" attribute also affects the prediction of SCC, with an accuracy drop of 0.036. These results confirm our hypothesis that metadata attributes significantly impact disease diagnosis and emphasize the importance of collecting valid and informative metadata to improve model diagnostic accuracy.

F. Reader Study

To evaluate the clinical usefulness of RemixFormer++, we conducted a comparative study of clinical diagnosis accuracy with 191 dermatologists across four categories of expertise: dermatology specialists (58), attending dermatologists (59), dermatology residents (49), and general clinical practitioners (25). We utilized an independent test set with 100 patients and 12 skin tumors for the reader study. Each physician was asked to randomly diagnose 20 patients with modality information provided in the order of C, CM, and CDM. Fig. 11a shows the performance comparison between dermatologists and RemixFormer++. The box plot indicates that the Top-1 accuracy of dermatologists at different levels significantly improved when patients' metadata and dermoscopy images were sequentially provided. This is consistent with the results of RemixFormer++. Notably, our algorithm outperformed the



(a) (b)

Fig. 11: Performance comparison of RemixFormer++ with dermatologists. (a) Comparative boxplot of diagnostic accuracy across different modalities of information among four levels of dermatologists and our RemixFormer++. The "+" symbol represents the mean accuracy. (b) ROC curves of the RemixFormer++ with three-modality information compared to four levels of dermatologists, where asterisks indicate the average values of each group of physicians. The predictive performance of the top 10 dermatology specialists is denoted by individual red dots. RemixFormer++ has achieved better performance than most dermatologists.

average dermatologists' performance across different modalities (indicated by crosses in Fig. 11a). Fig. 11b displays the receiver operating characteristic (ROC) curve of our algorithm (AUC=0.971) and the performance of dermatologists in CDM. RemixFormer++ achieved better performance than most dermatologists, which may suggest its potential clinical value.

VI. Discussion

From a cognitive perspective, the proposed RemixFormer++ achieves significant improvements in classification performance on multiple datasets due to effective cross-scale feature alignment. The challenges in clinical images and dermoscopy images are distinct, and the difference is rooted in the varying scales of the features. Therefore, different modalities must use separate strategies to generate biologically consistent features aligned with the same disease. The main technique problem is that the clinical images need to locate lesions from cluttered backgrounds, while dermoscopy images require the discernment of fine-grained textures at high resolutions. To address this issue, we drew inspiration from strategies employed by dermatologists in clinical practice to use separate models to deal with clinical images and dermoscopy images. More specifically, our clinical image branch employs a top-down strategy to accomplish the fusion of global and local features, and the LSM module provides auxiliary lesion information that significantly improves the overall performance of RemixFormer++ (Table III and X). Meanwhile, RemixFormer++ adopts a bottom-up two-level hierarchical architecture to capture both window-level and region-level features, and it achieves outstanding results by utilizing high-resolution window embeddings in HRWE. The window-level attention map in Fig. 8 illustrates that HRWE can focus/localize on important visual patterns, such as the pigment network (outlined in green), in dermoscopy images. Through the ablation studies shown in Table VII, Table X, and Table XII, it can be seen that LSM and HRWE contribute the most performance gains and well address the main technique challenges in the clinical and dermoscopy images, respectively.

Utilizing multi-modal data has been found to enhance significantly the diagnostic accuracy of skin tumors by both physicians and RemixFormer++, verified from the reader study. The model's performance using only clinical images is notably inferior to that of most dermatologists who use multi-modal data. Thus, our algorithm, which leverages the complementary benefits of multi-source data, is more suitable for the real clinical setting. We have observed that metadata plays an essential role in the diagnosis performance, as depicted in Fig. 11a. X-SkinTumor-12 metadata takes into account multiple factors (detailed in Table II) that are strongly associated with diagnosis such as disease progression, patients' subjective experiences, sun exposure (related to Melanoma), etc. Future work will quantitatively explore the impact of these factors on model performance. Additionally, we have found that dermoscopy images are more advantageous for senior physicians than junior physicians due to the latter's limited experience with dermoscopy. RemixFormer++ may offer potential assistance to primary physicians by better utilizing the dermoscopic channel of information. Going forward, we plan to apply this work to a multi-center clinical validation to further exploit the efficacy of RemixFormer++ in real-world clinical settings.

Deploying AI in clinical practice demands careful ethical oversight. We fully acknowledge the intricate debates surrounding morality, ethics, and legal frameworks, and are committed to addressing these issues with responsibility and balance. We adhere to stringent privacy protection protocols to ensure the security of patient data and to enhance the accuracy and transparency of AI. In collaboration with medical professionals, we ensure the prudent use and monitoring of model capabilities, with plans to evaluate the performance

of RemixFormer++ in real-world clinical settings. In essence, the transparency of AI processes, actively addressing patient concerns, and ensuring smooth doctor-patient communication are vital cornerstones of the healthcare journey.

VII. Conclusion

We have proposed an efficient multi-modal deep learning framework, RemixFormer++, that effectively combines multiple sources of data through three different branches. The clinical image branch incorporates local lesion features and global contextual information, while the dermoscopic image branch captures texture information and learns high-resolution image representations. The efficacy of the proposed clinical and dermoscopy branches has been independently validated on the public PAD-UFES-20, ISIC 2017, ISIC 2018, and ISIC 2019 datasets, demonstrating a competitive advantage. Through quantitative experiments, we have shown that LSM and HRWE can significantly improve classification performance. Combining enhanced clinical and dermoscopy branches with effective multi-modal fusion, our approach achieves a new record on the public multi-modal dataset Derm7pt, outperforming the previous best method by 5.3% in F1. We have further validated RemixFormer++ on a large-scale in-house X-SkinTumor-12 dataset, where it attains an overall classification accuracy of 92.6% for identifying 12 different types of skin tumors. Last our approach surpasses most of the 191 dermatologists on 100 test patient cases, demonstrating its excellent potential for clinical applications.

References

[1] A. Q. Garrido *et al.*, "Diagnosis of cutaneous melanoma: the gap between the knowledge of general practitioners and dermatologists in a brazilian population," *Journal of Cancer Education*, vol. 35, pp. 819–825, 2020.

[2] A. Memon *et al.*, "Changing epidemiology and age-specific incidence of cutaneous malignant melanoma in england: An analysis of the national cancer registration data by age, gender and anatomical site, 1981–2018," *The Lancet Regional Health - Europe*, vol. 2, p. 100024, 2021.

[3] B. K. Rao, J. Tan, and T. Bronsnick, *Moschella and Hurley's dermatology*. Jaypee Brothers Medical Pub, 2020.

[4] J. Dinnes *et al.*, "Visual inspection and dermoscopy, alone or in combination, for diagnosing keratinocyte skin cancers in adults." *The Cochrane database of systematic reviews*, vol. 12, p. CD011901, 2018.

[5] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 115–118, 2017.

[6] Z. Ge, S. Demyanov, R. Chakravorty, A. Bowling, and R. Garnavi, "Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images," in *MICCAI*, 2017, pp. 250–258.

[7] H. A. Haenssle *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*, vol. 31, no. 1, pp. 137–143, 2020.

[8] L. R. Soenksen, T. Kassis, S. T. Conover, B. Marti-Fuster, and M. L. Gray, "Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images," *Science Translational Medicine*, vol. 13, no. 581, p. eabb3652, 2021.

[9] P. Tang, X. Yan, Y. Nan, S. Xiang, S. Krammer, and T. Lasser, "Fusionm4net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification," *Medical Image Analysis*, vol. 76, no. 102307, pp. 1–13, 2022.

[10] P. Tschandl *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*, vol. 20, no. 7, pp. 938–947, 2019.

[11] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 538–546, 2019.

[12] A. G. Pacheco and R. A. Krohling, "The impact of patient clinical information on automated skin cancer detection," *Computers in biology and medicine*, vol. 116, p. 103545, 2020.

[13] L. Bi, D. D. Feng, M. Fulham, and J. Kim, "Multi-label classification of multi-modality skin lesion via hyper-connected convolutional neural network," *Pattern Recognition*, vol. 107, p. 107502, 2020.

[14] S. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines," *npj Digit. Medicine*, vol. 3, p. 136, 2020.

[15] J. Xu *et al.*, "Remixformer: A transformer model for precision skin tumor differential diagnosis via multi-modal imaging and non-imaging data," in *MICCAI*, 2022, pp. 624–633.

[16] A. Zakaria, T. A. Miclau, T. Maurer, K. S. Leslie, and E. Amerson, "Cost minimization analysis of a teledermatology triage system in a managed care setting," *JAMA Dermatology*, vol. 157, no. 1, pp. 52–58, 2021.

[17] N. Cowan, "What are the differences between long-term, short-term, and working memory?" *Progress in Brain Research*, vol. 169, no. 323–338, 2008.

[18] J. Yang, X. Sun, J. Liang, and P. L. Rosin, "Clinical skin lesion diagnosis using representations inspired by dermatologist criteria," in *CVPR*, 2018, pp. 1258–1266.

[19] I. González-Díaz, "Dermaknet: Incorporating the knowledge of dermatologists to convolutional neural networks for skin lesion diagnosis," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 547–559, 2019.

[20] K. Kamnitsas *et al.*, "Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation," *Medical Image Analysis*, vol. 36, pp. 61–78, 2017.

[21] N. Gessert *et al.*, "Skin lesion classification using cnns with patch-based attention and diagnosis-guided loss weighting," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 2, pp. 495–503, 2020.

[22] N. Tajbakhsh *et al.*, "Convolutional neural networks for medical image analysis: Full training or fine tuning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[23] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[24] W. Liu, Y. Wang, J. Tao, Y. Chi, L. Zhang, and X. sheng Hua, "Landmarks detection with anatomical constraints for total hip arthroplasty preoperative measurements," in *MICCAI*, 2020, pp. 670–679.

[25] M. A. Marchetti *et al.*, "Results of the 2016 international skin imaging collaboration international symposium on biomedical imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images," *Journal of the American Academy of Dermatology*, vol. 78, no. 2, pp. 270–277, 2018.

[26] Z. Wu *et al.*, "Studies on different CNN algorithms for face skin disease classification based on clinical images," *IEEE Access*, vol. 7, pp. 66 505–66 511, 2019.

[27] X. Zhao *et al.*, "The application of deep learning in the risk grading of skin tumors for patients using clinical images," *Journal of medical systems*, vol. 43, no. 8, pp. 283:1–283:7, 2019.

[28] Y. Xie, J. Zhang, Y. Xia, and C. Shen, "A mutual bootstrapping model for automated skin lesion segmentation and classification," *IEEE Transactions on Medical Imaging*, vol. 39, no. 7, pp. 2482–2493, 2020.

[29] M. A. Khan, M. Sharif, T. Akram, S. Kadry, and C. Hsu, "A two-stream deep neural network-based intelligent system for complex skin cancer types classification," *Int. J. Intell. Syst.*, vol. 37, no. 12, pp. 10 621–10 649, 2022.

[30] Z. Liu, R. Xiong, and T. Jiang, "Ci-net: Clinical-inspired network for automated skin lesion recognition," *IEEE Transactions on Medical Imaging*, vol. 42, no. 3, pp. 619–632, 2023.

[31] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin lesion classification using ensembles of multi-resolution efficientnets with meta data," *MethodsX*, vol. 7, p. 100864, 2020.

[32] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019, pp. 6105–6114.

[33] X. Fu, L. Bi, A. Kumar, M. Fulham, and J. Kim, "Graph-based intercategory and intermodality network for multilabel classification and melanoma diagnosis of skin lesions in dermoscopy and clinical images," *IEEE Transactions on Medical Imaging*, vol. 41, no. 11, pp. 3266–3277, 2022.

[34] Y. Liu *et al.*, "A deep learning system for differential diagnosis of skin diseases," *Nature Medicine*, vol. 26, pp. 900–908, 2020.
[35] J. Yang *et al.*, "Self-paced balance learning for clinical skin disease recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 8, pp. 2832–2846, 2019.
[36] K. Gupta, M. Krishnan, A. Narayanan, N. S. Narayan *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)*. IEEE, 2021, pp. 5262–5269.
[37] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021, pp. 10 012–10 022.
[38] J. Wang, X. Yu, and Y. Gao, "Feature fusion vision transformer for fine-grained visual categorization," *arXiv preprint arXiv:2107.02341*, 2021.
[39] A. Katharopoulos and F. Fleuret, "Processing megapixel images with deep attention-sampling models," in *ICML*, 2019, pp. 3282–3291.
[40] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.
[41] M. Caron *et al.*, "Emerging properties in self-supervised vision transformers," in *ICCV*, 2021, pp. 9650–9660.
[42] H. Zhang, J. Xue, and K. Dana, "Deep ten: Texture encoding network," in *CVPR*, 2017, pp. 708–717.
[43] C. Ou *et al.*, "A deep learning based multimodal fusion model for skin lesion diagnosis using smartphone collected clinical images and metadata," *Frontiers in Surgery*, vol. 9, 2022.
[44] J. Xu *et al.*, "Analysis of globule types in malignant melanoma," *Archives of dermatology*, vol. 145, no. 11, pp. 1245–1251, 2009.
[45] J. Yap, W. Yolland, and P. Tschandl, "Multimodal skin lesion classification using deep learning," *Experimental dermatology*, vol. 27, no. 11, pp. 1261–1267, 2018.
[46] Y. Wang *et al.*, "Adversarial multimodal fusion with attention mechanism for skin lesion classification using clinical and dermoscopic images," *Medical Image Analysis*, vol. 81, p. 102535, 2022.
[47] L. M. de Lima and R. A. Krohling, "Exploring advances in transformers and cnn for skin lesion diagnosis on small datasets," in *Intelligent Systems*, 2022, pp. 282–296.
[48] Y.-J. Zhou *et al.*, "A novel multi-task model imitating dermatologists for accurate differential diagnosis of skin diseases in clinical images," in *MICCAI*, 2023, pp. 202–212.
[49] C. Barata, M. E. Celebi, and J. S. Marques, "Explainable skin lesion diagnosis using taxonomies," *Pattern Recognition*, vol. 110, p. 107413, 2021.
[50] L. Wei, K. Ding, and H. Hu, "Automatic skin cancer detection in dermoscopy images based on ensemble lightweight deep learning network," *IEEE Access*, vol. 8, pp. 99 633–99 647, 2020.