Indepth Integration of Multi-granularity Features from Dual-modal for Disease Classification

Yeli Wu[1], Xiaocai Zhang[2], Weiwen Wu[1], Haiteng Jiang[3], Chao An[4(✉)], and Jianjia Zhang[1(✉)]

[1] School of Biomedical Engineering, Shenzhen Campus of Sun Yat-sen University, Shenzhen, 518107 China, wuyli29@mail2.sysu.edu.cn wuweiw7@mail.sysu.edu.cn zhangjj225@mail.sysu.edu.cn

[2] Faculty of Engineering and Information Technology, The University of Melbourne, Parkville, VIC 3010, Australia, xiaocai.zhang@unimelb.edu.au

[3] Advanced Brain Cognition and Disease Laboratory (ABCD Lab), School of Brain Science and Brain Medicine, Zhejiang University, 310058, Hangzhou, China, h.jiang@zju.edu.cn

[4] Chinese PLA General Hospital, 100853, Beijing, China, anchao-1983@163.com

Abstract. Multi-granularity features can be extracted from multiple modal medical images and how to effectively analyze these features is a challenging and critical issue for computer-aided diagnosis (CAD). However, most existing multi-modal classification methods have not fully explored the interactions among the intra- and inter-granularity features across multi-modal. To address this limitation, we propose a novel Indepth Integration of Multi-Granularity Features Network (IIMGF-Net) for a typical multi-modal task, i.e., a dual-modal based CAD. Specifically, the proposed IIMGF-Net consists of two types of key modules, i.e., Cross-Modal Intra-Granularity Fusion (CMIGF) and Multi-Granularity Collaboration (MGC). The CMIGF module enhances the attentive interactions between the same granularity features from dual-modals and derive an integrated representation at each granularity. Based on these representations, the MGC module captures inter-granularity interactions among the resulting representations of CMIGF through the coarse-to-fine and fine-to-coarse collaborative learning mechanism.Extensive experiments on two dual-modal datasets validate the effectiveness of the proposed method, demonstrating its superiority in dual-modal CAD tasks by integrating multi-granularity information.

Keywords: Multi-granularity · Dual-modal medical images · Computer-aided Diagnosis · Feature Fusion · Classification.

1 Introduction

In recent years, with the rapid advancement of medical technology, there has been a growing interest in leveraging multi-modal medical images for computer-aided diagnosis (CAD) [17,21,22,25]. As a typical case in practical scenarios, dual-modal images are usually collected for analysis. With the latest deep learning

techniques, multi-granularity features can be extracted from each image modal to reflect the coarse- and fine-grained properties of the images. A crucial aspect of dual-modal based CAD is to effective analyze these multi-granularity features, and this involves the fusion and collaboration of multi-granularity features from all the modals.

The necessity and importance of analyzing multi-granularity features can be demonstrated by the following two specific medical applications, i.e., skin cancer diagnosis and lymph node metastasis prediction. In the field of skin cancer diagnosis, coarse-grained information captures global lesion characteristics [22], while fine-grained details provide critical insight into pigment distribution, boundary sharpness, and texture characteristics [11], and all of which are essential for accurate diagnosis. Similarly, in the field of lymph node metastasis prediction [2], multi-granularity features enable the extraction of richer structural and functional information [2,23], thereby enhancing predictive accuracy.

Despite the significance and potential advantages, the existing multi-modal image classification methods, which can be primarily categorized into early- [6,15], middle- [9,14,22,25], and late-fusion [19,20] methods, still suffer notable shortcomings. The early-fusion methods [6,15] usually preprocess the raw multi-modal data or concatenate them directly, preventing the indepth interactions between the image modal. Middle-fusion methods alleviate this issue by performing fusion at the feature level.

Nevertheless, the interactions among the intra- and inter-granularity features across multi-modal have not been fully explored. For example, [14] and [22] employ cross-attention to fuse multi-modal features at a single granularity, but they do not explore the feature interactions among multiple granularities.

In contrast, [9] and [25] utilize fusion modules to enable feature interactions at different levels, but they still fail to capture the feature interactions across multiple granularities. Similarly, late-fusion methods such as [19] and [20] usually apply simple averaging or concatenating operations at the final classification stage, disregarding the intrinsic dependencies between multi-granularity features and constraining the model's ability to leverage multi-modal information effectively. Modeling the inter-granularity interactions is crucial for capturing hierarchical dependencies between coarse- and fine-grained features. The limitations of the existing methods hinder the optimal integration of complementary and discriminative information, leading to suboptimal performance in multi-modal image classification.

To address these limitations, we propose an Indepth Integration of Multi-Granulairty Features Network (IIMGF-Net), which comprises two core modules, i.e., the Cross-Modal Intra-Granularity Fusion (CMIGF) and the Multi-Granularity Collaboration (MGC). Specifically, the CMIGF module is designed to tackle the issue of intra-granularity feature integration. At each granularity level, it adaptively fuse the dual-modal features with attentive Mamba [5] to derive an unified representation. Meanwhile, inspired by [26], the MGC module focuses on enhancing inter-granularity feature interactions by employing a coarse-to-fine and fine-to-coarse mechanism. It enables bidirectional collabora-

tion between multi-granularity features, establishing hierarchical dependencies across different granularity features. By integrating intra- and inter-granularity features more comprehensively, the proposed model ensures richer feature extraction from dual-modal images, ultimately enhancing the classification performance.

The key contributions of this paper can be summarized as follows.

1. We propose a novel IIMGF-Net, which can achieve indepth integration of multi-granularity features from dual-modal medical images to improve the accuracy of disease diagnosis.
2. We explore a Cross-Modal Intra-Granularity Fusion module and the Multi-Granularity Collaboration module, which enable intra-granularity feature fusion and inter-granularity feature collaboration, respectively.
3. We conduct comprehensive experiments to validate the effectiveness of the proposed IIMGF-Net. Compared to the state-of-the-art methods, the proposed method achieves the highest classification performance.

2 Methodology

2.1 Overall Architecture

We propose the Indepth Integration of Multi-Granularity Features Network (IIMGF-Net), which consists of two primary stages, as illustrated in Fig. 1(a). The first stage comprises L stacked encoders, each employing a dual-stream VSS module followed by a down-sampling module as proposed in [13], as the backbone architecture to extract multi-granularity features from the dual-modal inputs. Specifically, the raw dual-modal images, denoted as $\boldsymbol{x}_{raw}$ and $\boldsymbol{y}_{raw}$, are preprocessed by convolutional layer and ReLu activation function to obtain $\boldsymbol{x}_0, \boldsymbol{y}_0 \in \mathbb{R}^{H_0 \times W_0 \times C_0}$, which are then fed into the first encoder. Within the stacked encoders, multi-granularity features are iteratively extracted as follows:

$$\boldsymbol{x}_l = \mathrm{VSS}(\boldsymbol{x}_{l-1}), \quad \boldsymbol{y}_l = \mathrm{VSS}(\boldsymbol{y}_{l-1}), l \in [1, \cdots, L] \tag{1}$$

where $\boldsymbol{x}_l, \boldsymbol{y}_l \in \mathbb{R}^{\frac{H_0}{2^l} \times \frac{W_0}{2^l} \times 2^l C_0}$.

To enhance the intra-granularity features interactions, a Cross-Modal Intra-Granularity Fusion (CMIGF) module in each encoder adaptively fuses $\boldsymbol{x}_l$ and $\boldsymbol{y}_l$ and generates $\boldsymbol{f}_x^l$ and $\boldsymbol{f}_y^l$. $\boldsymbol{f}_x^l$ and $\boldsymbol{f}_y^l$ are concatenated to form $\boldsymbol{Z}_l$ as a joint representation at this granularity, and it will be fed into the Multi-Granularity Collaboration (MGC) module at the second stage for inter-granularity interaction analysis. At the same time, $\boldsymbol{f}_x^l$ and $\boldsymbol{f}_y^l$ will be added to $\boldsymbol{x}_{l+1}$ and $\boldsymbol{y}_{l+1}$ as the input of the CMIGF in the next encoder.

The outputs $\boldsymbol{x}_L$ and $\boldsymbol{y}_L$ from the last encoder are added to $\boldsymbol{f}_x^L$ and $\boldsymbol{f}_y^L$ respectively, extracted by the CMIGF module in the last encoder. The resulting features and the output of the MGC module are then processed through the global average pooling (GAP) layer. The pooled features are subsequently concatenated and fed into a fully connected layer for classification.

The details of the two key modules of CMIGF and MGC are introduced below.

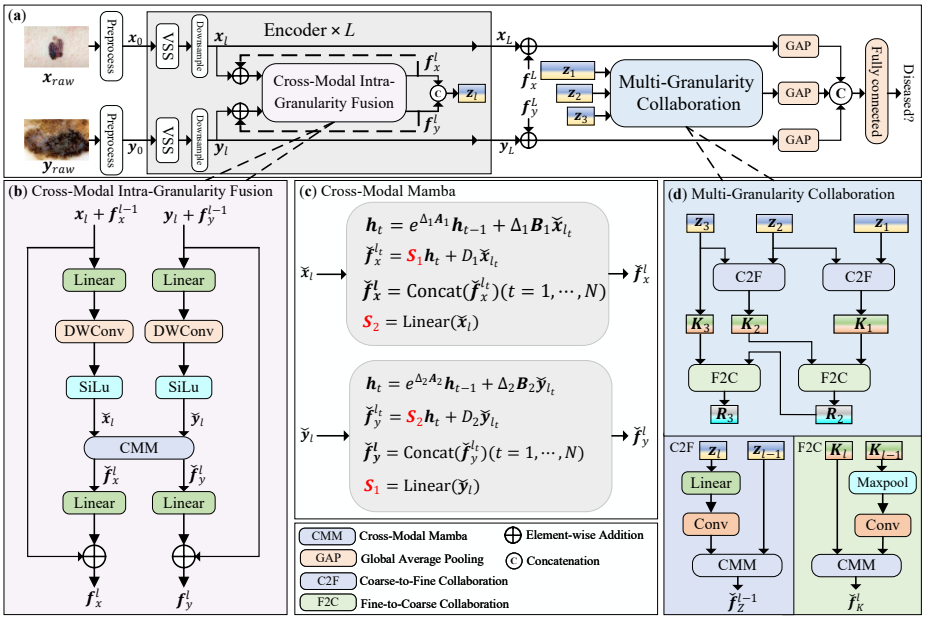

Fig. 1. Illustration of our proposed network. (a) The overall architecture. In the first stage, input data are processed through stacked encoders, and their outputs are concatenated and fed into the Multi-Granularity Collaboration (MGC) in the second stage. The extracted features from both stages undergo Global Average Pooling (GAP), are concatenated, and then passed through a fully connected layer for final prediction. (b) The Cross-Modal Intra-Granularity Fusion (CMIGF) module. (c) The Cross-Modal Mamba (CMM) module. (d) The Multi-Granularity Collaboration (MGC) module.

2.2 Cross-Modal Intra-Granularity Fusion

In order to adaptively fuse the features at the same granularity and extract the complementary information from them, a Cross-modal Intra-granularity Fusion (CMIGF) module (illustrated in Fig. 1(b)) is utilized. With input $\boldsymbol{x}_l, \boldsymbol{y}_l$, the CMIGF module can be formulated as follows:

$$\begin{aligned}
\check{\boldsymbol{x}}_l, \check{\boldsymbol{y}}_l &= \big(\mathrm{SiLu}(\mathrm{DWConv}(\mathrm{Linear}(\boldsymbol{x}_l + \boldsymbol{f}_x^{l-1}))), \\
&\qquad \mathrm{SiLu}(\mathrm{DWConv}(\mathrm{Linear}(\boldsymbol{y}_l + \boldsymbol{f}_y^{l-1})))\big), \\
\check{\boldsymbol{f}}_x^l, \check{\boldsymbol{f}}_y^l &= \mathrm{CMM}(\check{\boldsymbol{x}}_l, \check{\boldsymbol{y}}_l), \\
\boldsymbol{f}_x^l, \boldsymbol{f}_y^l &= \big(\boldsymbol{x}_l + \boldsymbol{f}_x^{l-1} + \mathrm{Linear}(\check{\boldsymbol{f}}_x^l), \quad \boldsymbol{y}_l + \boldsymbol{f}_y^{l-1} + \mathrm{Linear}(\check{\boldsymbol{f}}_y^l)\big).
\end{aligned} \tag{2}$$

The inputs are sequentially processed through a linear mapping, depth-wise convolution (DWConv) and SiLu activation function, generating intermediate representations $\check{\boldsymbol{x}}_l, \check{\boldsymbol{y}}_l \in \mathbb{R}^{N_l \times 2lC_0}$ where $N_l = \frac{H_0}{2^l} \times \frac{W_0}{2^l}$. These representations are then fed into the Cross-Modal Mamba (CMM) module (described in detail in

the next section) to generate the interactive outputs $\check{\boldsymbol{f}}_x^l, \check{\boldsymbol{f}}_y^l \in \mathbb{R}^{N_l \times 2lC_0}$. Subsequently, a linear mapping restores the original feature dimension $2lC_0$, while the spatial dimension is recovered by reshaping N_l back into $\frac{H_0}{2^l} \times \frac{W_0}{2^l}$. Finally, a residual connection ensures that the outputs $\boldsymbol{f}_x^l, \boldsymbol{f}_y^l$ retain the original input information. $\boldsymbol{f}_x^l$ and $\boldsymbol{f}_y^l$ are concatenated to form $\boldsymbol{Z}_l$ as a joint representation of them at this granularity, and they will be fed back into the CMIGF in the next encoder, with $\boldsymbol{f}_x^0, \boldsymbol{f}_y^0 = \boldsymbol{0}$.

2.3 Cross-Modal Mamba

The Cross-Modal Mamba (CMM) module, illustrated in Fig. 1(c), is designed to enable adaptive cross-modal interactions and facilitate information communication between the two modalities under the Mamba framework [5]. With inputs $\check{\boldsymbol{x}}_l, \check{\boldsymbol{y}}_l \in \mathbb{R}^{N_l \times 2lC_0}$, $\check{\boldsymbol{f}}_x^l$ can be obtained as follows:

$$\begin{aligned} \boldsymbol{h}_t &= e^{\Delta_1 \boldsymbol{A}_1} \boldsymbol{h}_{t-1} + \Delta_1 \boldsymbol{B}_1 \check{\boldsymbol{x}}_{l_t}, \\ \check{\boldsymbol{f}}_x^{l_t} &= \boldsymbol{S}_1 \boldsymbol{h}_t + D_1 \check{\boldsymbol{x}}_{l_t}, \\ \check{\boldsymbol{f}}_x^l &= \text{Concat}(\check{\boldsymbol{f}}_x^{l_t})(t \in [1, \cdots, N]), \end{aligned} \tag{3}$$

where $\check{\boldsymbol{x}}_{l_t} \in \mathbb{R}^{1 \times 2lC_0}$ represents represents the input at the token level, while $\check{\boldsymbol{f}}_x^l \in \mathbb{R}^{N_l \times 2lC_0}$ denotes the final output. The term $\Delta_1 \in \mathbb{R}$ is a timescale parameter used to discretize $\boldsymbol{A}_1$ and $\boldsymbol{B}_1$. The learnable parameters $\boldsymbol{A}_1 \in \mathbb{R}^{d \times d}$ and $D_1 \in \mathbb{R}$ are randomly initialized, whereas $\boldsymbol{B}_1 \in \mathbb{R}^{d \times 1}$ and $\Delta_1 \in \mathbb{R}$ are specifically associated with $\check{\boldsymbol{x}}_l$. $\boldsymbol{S}_1 \in \mathbb{R}^{1 \times d}$ is obtained by $\boldsymbol{S}_1 = \text{Linear}(\check{\boldsymbol{y}}_l)$, where d denotes the hidden dimension. Through this sequential scanning process, the hidden state $\boldsymbol{h}_t \in \mathbb{R}^{d \times 2lC_0}$ progressively accumulates contextual information, capturing long-range dependencies between tokens. $\check{\boldsymbol{f}}_y^l$ can be obtained similarly by switching x and y in the equations above. In this case, the CMM module produces the outputs $\check{\boldsymbol{f}}_x^l$ and $\check{\boldsymbol{f}}_y^l$.

2.4 Multi-Granularity Collaboration

Inspired by [26], the Multi-Granularity Collaboration (MGC) module illustrated in Fig. 1(d) is incorporated into the proposed model to explore inter-granularity feature interaction. As introduced above, multi-granularity features $\boldsymbol{Z} = \{\boldsymbol{Z}_l\}_{l=1}^L$ ($L = 3$ in this study) can be generated by the CMIGF module, i.e., $\boldsymbol{Z}_l = \text{Concat}(\boldsymbol{f}_x^l, \boldsymbol{f}_y^l) \in \mathbb{R}^{\frac{H_0}{2^l} \times \frac{W_0}{2^l} \times 4lC_0}$. Different granularity features may present specific image properties. For example, the fine-grained features (with smaller l) may mainly present detailed lesion boundary prediction, while coarse-grained features (with larger l) capture broader contextual information. These multi-granularity features are probably closely correlated to each other and should be jointly analyzed for disease diagnosis. To enable comprehensive cross-granularity interactions between these features, a bidirectional coarse-to-fine and fine-to-coarse

collaborative learning mechanism is developed in MGC module, as introduced below.

Coarse-to-Fine Collaboration (C2F). With $\boldsymbol{Z}_l$ as the fine-grained feature and $\boldsymbol{Z}_{l+1}$ as the coarse-grained feature, the key idea of this C2F module is to align the coarse-grained feature $\boldsymbol{Z}_{l+1}$ with $\boldsymbol{Z}_l$ first by enlarging it and then perform attentive cross-granularity interactions between them, finally deriving a joint representation $\boldsymbol{K}_l$. Specifically, to align $\boldsymbol{Z}_{l+1}$ with $\boldsymbol{Z}_l$, we first apply a linear transformation followed by a convolutional layer for dimension expansion. The coarse-to-fine collaboration is then achieved via the CMM module.

$$\begin{aligned} \hat{\boldsymbol{Z}}_{l+1} &= \mathrm{Conv}((\mathrm{Linear}(\boldsymbol{Z}_{l+1}))), \\ \boldsymbol{K}_l &= \mathrm{CMM}(\boldsymbol{Z}_l, \hat{\boldsymbol{Z}}_{l+1}) \end{aligned} \tag{4}$$

Fine-to-Coarse Collaboration (F2C). Converse to C2F, F2C aims to shrink the fine-grained feature and then perform attentive cross-granularity interactions between the shrunk feature and the coarse-grained feature. Specifically, with the resulting feature $\boldsymbol{K}_1$ from the C2F module as the fine-grained feature and $\boldsymbol{K}_2$ as the coarse-grained feature, $\boldsymbol{K}_1$ is downsampled via max-pooling, followed by a convolutional layer to align its dimensions with $\boldsymbol{K}_2$. Then the fine-to-coarse cross-granularity interactions are performed with the CMM module. The process can be formulated as:

$$\begin{aligned} \hat{\boldsymbol{K}}_1 &= \mathrm{Conv}\left(\mathrm{MaxPooling}\left(\boldsymbol{K}_1\right)\right) \\ \boldsymbol{R}_2 &= \mathrm{CMM}\left(\boldsymbol{K}_2, \hat{\boldsymbol{K}}_1\right) \end{aligned} \tag{5}$$

In the following, $\boldsymbol{R}_2$ is considered as the fine-grained feature and $\boldsymbol{K}_3$ as the coarse-grained feature for the same process as described above to obtain the final refined feature $\boldsymbol{R}_3$. By doing this, the feature $\boldsymbol{R}_3$ are enhanced by incorporating both coarse- and fine-grained lesion clues for diagnosis.

3 Experiments and Results

3.1 Datasets

In this study, two tumor datasets, including one public dataset and one private dataset, are used in the evaluation.

1. The publicly available Derm7pt dataset [12] comprises 413 training samples, 203 validation samples, and 395 test samples. Each sample has a dermoscopy and a clinical image. It includes eight labels: a diagnostic label and seven 7-point checklist labels, i.e., pigment network (PN), blue-whitish veil (BWV), vascular structures (VS), pigmentation (PIG), streaks (STR), dots and globules (DaG), and regression structures (RS). The task is to predict the eight labels for each sample. The data splits provided by the original study's [12] are used.

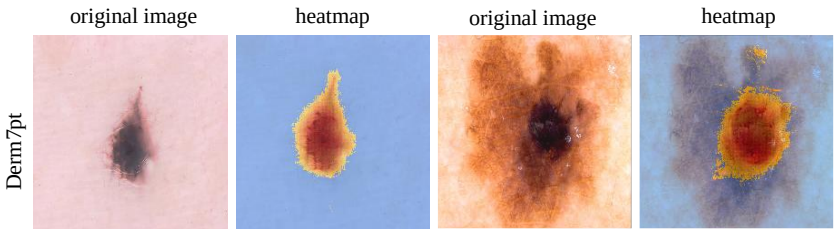

Fig. 2. Integrated Gradients visualization on the Derm7pt datasets.

2. DECT-LNM Dataset: The private dataset consists of a total of 160 samples, each containing six dual-energy CT (DECT) images: three images obtained at 100 keV and the rest at 150 keV. These images containing adenocarcinoma were cropped from the raw DECT images. Subsequently, each cropped image is normalized and resized to 224 × 224 pixels. Among these samples, 75 are positive (lymph node metastases) and 85 are negative (no metastases). Five-fold cross-validation is performed.

3.2 Implement details

We employ Vmamba [13] as the pretrained backbone, and the raw inputs are $\boldsymbol{x}_{raw}, \boldsymbol{y}_{raw} \in \mathbb{R}^{224\times224\times3}$, while the final outputs are $\boldsymbol{x}_L, \boldsymbol{y}_L \in \mathbb{R}^{7\times7\times768}$, with the encoder depth set to $L = 3$. Data augmentation includes random flips, shifts, rotations, and brightness-contrast adjustments. The model is trained using AdamW optimizer with a learning rate of 1e-4, weight decay of 1e-4, and batch size of 32. All the experiments are performed on a single NVIDIA RTX A6000 with 48 GB memory. The code is publicly available at `https://github.com/seuzjj/IIMGF-Net`

3.3 Results

Twelve state-of-the-art methods are involved in the experimental study, including Rest [24], Xcit [1], Levit [4], Mambavision [7], Efficientnet [18], RegNet [16], ResNet [8], and Xception [3] as early fusion models; HiFuse [10], CRD-Net [14], and TFormer [25] as middle fusion models; and the late fusion model of FM4Net [19]. As shown in Table 1, 'E' denotes early fusion, 'M' denotes middle fusion, and 'L' denotes late fusion.

On Derm7pt Dataset Table 1 presents the accuracy (ACC), Area Under the Curve (AUC), sensitivity (SEN), and specificity (SPE) averaged over the eight classification tasks. As can be seen, the proposed IIMGF-Net achieves the highest ACC of 77.0% among all the competing methods, outperforming the state-of-the-art multi-modal methods CRO-Net, FM4Net and TFormer by 5.4%, 3.2%, and 2.1%, respectively. The proposed method consistently achieves the best results

Table 1. The classification performance (%) of different methods on Dern7pt and DECT-LNM data sets.

Method	Derm7pt				DECT-LNM				Category
	ACC	AUC	SEN	SPE	ACC	AUC	SEN	SPE	
HiFuse[10]	66.0	72.3	49.7	72.7	82.0±3.8	83.9±3.1	79.2±8.8	83.5±8.4	M
Rest[24]	67.6	74.6	54.0	71.8	80.7±1.5	85.9±4.8	90.6±6.9	72.2±6.1	E
Xcit[1]	67.9	72.4	51.5	73.2	82.0±1.8	85.0±2.7	93.0±9.5	72.2±6.0	E
Levit[4]	68.8	75.2	55.3	73.9	84.0±3.7	85.4±4.9	86.4±7.5	80.1±9.2	E
Mambavision[7]	68.9	74.9	56.2	77.4	82.7±4.4	86.5±5.3	85.6±6.6	78.9±12.8	E
Efficientnet[18]	69.1	76.9	55.5	76.4	84.7±4.5	86.6±3.1	92.3±7.7	78.5±10.2	E
RegNet[16]	69.2	76.6	54.0	76.0	84.0±1.5	86.7±4.7	82.4±11.5	84.6±9.5	E
ResNet[8]	69.6	76.8	55.6	76.4	84.0±4.4	85.7±4.6	88.4±4.2	79.8±7.9	E
Xception[3]	70.6	78.4	55.1	76.2	85.3±6.5	86.5±7.5	94.7±7.9	76.7±10.7	E
CRD-Net[14]	71.6	80.4	58.7	78.7	87.3±4.4	89.7±6.7	96.5±5.3	80.2±8.6	M
FM4Net [19]	73.8	83.0	61.8	80.4	86.7±4.1	90.2±4.4	93.2±4.4	81.0±8.8	L
TFormer[25]	74.9	84.8	63.5	81.3	86.0±2.8	89.1±4.6	88.7±7.4	83.6±5.3	M
IIMGF-Net(proposed)	**77.0**	**85.8**	**65.3**	**82.7**	**91.3**±5.1	**94.7**±4.9	**95.2**±7.9	**87.1**±4.5	M

*E: Early fusion, M: Middle fusion, L: Late fusion.

Table 2. The abation experiment (%) of different methods on Derm7pt and DECT-LNM data sets.

Module	Derm7pt				DECT-LNM			
	ACC	AUC	SEN	SPE	ACC	AUC	SEN	SPE
without MGC	76.1	**85.8**	63.4	81.2	88.7±3.0	91.7±3.2	92.6±4.5	85.0±6.7
without CMIGF	75.7	84.5	63.2	81.1	87.3±2.8	92.0±3.8	86.7±11.1	86.9±7.4
IIMGF-Net(proposed)	**77.0**	**85.8**	**65.3**	**82.7**	**91.3**±5.1	**94.7**±4.9	**95.2**±7.9	**87.1**±4.5

in AUC, SEN, and SPE. Moreover, as shown by the attention score in Fig. 2, the model reasonably focus on the lesion areas to reach the diagnosis conclusion. These results highlight the superior performance of the proposed IIMGF-Net compared to the state-of-the-art methods.

On DECT-LNM Dataset Table 1 presents the evaluation results with multiple metrics on DECT-LNM dataset. As can be seen, the proposed IIMGF-Net achieves the highest ACC of 91.3% among all the competing methods, outperforming the state-of-the-art multi-modal methods TFormer, FM4Net and CRO-Net by 5.3%, 4.6% and 4.0%, respectively. The proposed method consistently achieves the best results in AUC, SEN, and SPE, further confirming its effectiveness.

Ablation Studies Ablation studies are conducted to verify the efficacy of the CMIGF and MGC modules in the proposed method, and the results are presented in Table 2. As can be seen, on the Derm7pt dataset, removing MGC or CMIGF reduces accuracy by 0.9% and 1.3%, respectively. On the DECT-LNM dataset, the accuracy of the proposed method will be decreased by 2.6% and 4.0% if without the MGC or CMIGF module, respectively. This highlights

the importance of exploring both intra-granularity and inter-granularity features interactions in multi-modal-based diagnosis tasks.

4 Conclusion

In sum, this paper proposes an IIMGF-Net to explore the interactions among the intra- and inter-granularity features across dual-modals. Comprehensive experiments on the Derm7pt and DECT-LNM datasets consistently validate the effectiveness of the proposed method, achieving state-of-the-art classification performance.

Acknowledgments. This work was supported in part by Shenzhen Medical Research Fund under Grant B2402030; in part by the National Natural Science Foundation of China under Grant 62471501, Grant 62471502; in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515012278; in part by Shenzhen Science and Technology Program under Grant JCYJ20240813151026034 and JCYJ 20220818102414031; in part by the Open Research Fund of the State Key Laboratory of Brain-Machine Intelligence, Zhejiang University, under Grant BMI2400019.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ali, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., Laptev, I., Neverova, N., Synnaeve, G., Verbeek, J., et al.: Xcit: Cross-covariance image transformers. Advances in neural information processing systems **34**, 20014–20027 (2021)
2. An, C., Li, D., Li, S., Li, W., Tong, T., Liu, L., Jiang, D., Jiang, L., Ruan, G., Hai, N., et al.: Deep learning radiomics of dual-energy computed tomography for predicting lymph node metastases of pancreatic ductal adenocarcinoma. European journal of nuclear medicine and molecular imaging pp. 1–13 (2022)
3. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017)
4. Graham, B., El-Nouby, A., Touvron, H., Stock, P., Joulin, A., Jégou, H., Douze, M.: Levit: a vision transformer in convnet's clothing for faster inference. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12259–12269 (2021)
5. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
6. Hashim, S., Ali, M., Nandakumar, K., Yaqub, M.: SubOmiEmbed: self-supervised representation learning of multi-omics data for cancer type classification. In: 2022 10th International Conference on Bioinformatics and Computational Biology (ICBCB). pp. 66–72. IEEE (2022)
7. Hatamizadeh, A., Kautz, J.: Mambavision: A hybrid mamba-transformer vision backbone. arXiv preprint arXiv:2407.08083 (2024)

8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. He, X., Wang, Y., Zhao, S., Chen, X.: Co-attention fusion network for multimodal skin cancer diagnosis. Pattern Recognition **133**, 108990 (2023)
10. Huo, X., Sun, G., Tian, S., Wang, Y., Yu, L., Long, J., Zhang, W., Li, A.: HiFuse: Hierarchical multi-scale feature fusion network for medical image classification. Biomedical Signal Processing and Control **87**, 105534 (2024)
11. Kang, K., Xu, J., An, S., Chen, J., Wang, R., Wang, H.: Multimodal hybrid approach for fine-grained classification of diverse dermatological conditions. In: 2024 2nd International Conference on Intelligent Perception and Computer Vision (CIPCV). pp. 157–164. IEEE (2024)
12. Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G.: Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE journal of biomedical and health informatics **23**(2), 538–546 (2018)
13. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., Liu, Y.: Vmamba: Visual state space model. Advances in neural information processing systems **37**, 103031–103063 (2025)
14. Liu, Z., Hu, Y., Qiu, Z., Niu, Y., Zhou, D., Li, X., Shen, J., Jiang, H., Li, H., Liu, J.: Cross-modal attention network for retinal disease classification based on multi-modal images. Biomedical Optics Express **15**(6), 3699–3714 (2024)
15. Odusami, M., Maskeliūnas, R., Damaševičius, R., Misra, S.: Explainable deep-learning-based diagnosis of alzheimer's disease using multimodal input fusion of pet and mri images. Journal of Medical and Biological Engineering **43**(3), 291–302 (2023)
16. Radosavovic, I., Kosaraju, R.P., Girshick, R., He, K., Dollár, P.: Designing network design spaces. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10428–10436 (2020)
17. Shu, C., Yu, L., Tian, S., Shi, X.: MSMA: A multi-stage and multi-attention algorithm for the classification of multimodal skin lesions. Biomedical Signal Processing and Control **93**, 106180 (2024)
18. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
19. Tang, P., Yan, X., Nan, Y., Xiang, S., Krammer, S., Lasser, T.: FusionM4Net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification. Medical Image Analysis **76**, 102307 (2022)
20. Wang, S., Yin, Y., Wang, D., Wang, Y., Jin, Y.: Interpretability-based multimodal convolutional neural networks for skin lesion diagnosis. IEEE transactions on cybernetics **52**(12), 12623–12637 (2021)
21. Xiao, C., Zhu, A., Xia, C., Qiu, Z., Liu, Y., Zhao, C., Ren, W., Wang, L., Dong, L., Wang, T., et al.: Attention-guided learning with feature reconstruction for skin lesion diagnosis using clinical and ultrasound images. IEEE Transactions on Medical Imaging (2024)
22. Xu, J., Huang, K., Zhong, L., Gao, Y., Sun, K., Liu, W., Zhou, Y., Guo, W., Guo, Y., Zou, Y., et al.: Remixformer++: A multi-modal transformer model for precision skin tumor differential diagnosis with memory-efficient attention. IEEE Transactions on Medical Imaging (2024)
23. Yang, A., Xu, L., Qin, N., Huang, D., Liu, Z., Shu, J.: MFU-Net: a deep multimodal fusion network for breast cancer segmentation with dual-layer spectral detector ct. Applied Intelligence **54**(5), 3808–3824 (2024)

24. Zhang, Q., Yang, Y.B.: Rest: An efficient transformer for visual recognition. Advances in neural information processing systems **34**, 15475–15485 (2021)
25. Zhang, Y., Xie, F., Chen, J.: TFormer: A throughout fusion transformer for multimodal skin lesion diagnosis. Computers in Biology and Medicine **157**, 106712 (2023)
26. Zhou, Z., Zhou, J., Qian, W., Tang, S., Chang, X., Guo, D.: Dense audio-visual event localization under cross-modal consistency and multi-temporal granularity collaboration. arXiv preprint arXiv:2412.12628 (2024)