I2M2Net: Inter/Intra-modal Feature Masking Self-distillation for Incomplete Multimodal Skin Lesion Diagnosis

Ke Wang[1], Linwei Qiu[1], Yilan Zhang[2], and Fengying Xie[1]*

[1] Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China
[2] King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia

Abstract. Multimodal learning has demonstrated promising advantages over single-modal approaches in the diagnosis of skin lesions. However, these methods often suffer from significant accuracy degradation when encountering missing modalities, hindering their clinical deployment. In this paper, we introduce a novel and effective framework, I2M2Net, for incomplete multimodal learning, focusing on adaptively and progressively mining knowledge about modal feature-aware combinations. Specifically, one branch conducts normal classification using the original complete multimodal features extracted by heterogeneous modal encoders, while another branch shares the same structures and weights, designed to perform self-distillation with masked modality combinations. These combinations are imposed on the complete features using two masking strategies simultaneously: 1) random dropout of modality (*i.e.* inter-modal feature masking) to simulate different missing modality combinations and foster combination-invariant dependencies, and 2) randomly mask patches of the remaining modal features (*i.e.* intra-modal feature masking) to promote combination-specific representations. Additionally, we design a combination-based curriculum learning (CCL) algorithm to identify weak combinations and progressively guide our network to facilitate incomplete modality learning on challenging combinations. This is achieved by adaptively adjusting the probabilities of masking based on the consistency between the complete combination and other combinations. Experimental results on the multimodal skin disease dataset Derm7pt demonstrate that our method outperforms other state-of-the-art approaches.

Keywords: Missing modality · Multi-modal learning · Skin lesion diagnosis.

1 Introduction

Skin cancer is a prevalent form of cancer, and early and precise diagnosis can greatly enhance the cure and survival rates, particularly in cases of melanoma [2]. In the clinical diagnosis, dermatologists typically take into account several

factors, including clinical images, dermoscopic images, and reference metadata (*e.g.* patient information and medical history), to get final diagnosis. Inspired by this, current research on automatic skin lesion diagnosis has explored various methods that leverage comprehensive multimodal information, rather than relying solely on a single modality[24,25,22], significantly enhancing classification performance[19,20,23]. However, issues such as data corruption, data acquisition failures and unclear associations of multimodal data, often result in missing clinical information. Most existing multimodal models may struggle to handle incomplete modal information, leading to a notable reduction in model classification accuracy[21]. Thus, it is necessary to develop multimodal models that are robust to incomplete modal data.

A typical solution is to use generative networks to directly synthesize missing modalities in the input data [3,15]. However, generative networks are often challenging to train, and the quality of the generated data significantly impacts the model's final performance[27]. As an alternative, methods based on knowledge distillation [4,6,14] or matrix completion [11] have been proposed to recover missing modality features. While these approaches have shown significant improvements, they necessitate training and deploying specific models for each combination of missing modalities, resulting in high time and space costs which are impractical for real-world applications. Current researches primarily focus on training a unified model to handle all potential combinations of missing modalities. For example, TATE [21] has developed a tag encoding module to address various missing modality scenarios and incorporated a new common space projection module for learning joint representations. MMIN [26] has employed a cross-modal imagination module to learn robust joint multimodal representations. Additionally, Lee *et al.* [10] have designed modality-missing-aware prompts plugged into multimodal transformers to handle general modality missing scenarios. Although the efficiency has been improved, these unified methods primarily concentrate on unified invariant features of various missing modality combinations while overlooking the mining of information in specific combination (especially the complementary modality-specific information). Consequently, unified models generally exhibit inferior performance compared to the customized models. Considering the diverse types and similar clinical presentations of skin lesions, leveraging specific details from multimodal information can aid in identifying similar categories.

To this end, we have designed a novel and effective inter/intra-modal feature masking self-distillation framework, where one branch of the self-distillation takes the complete modal features while the other branch utilizes the masked modality combination. Initially, masking occurs between modalities, which simulates various scenarios of modal absence. This inter-modal mask aims to facilitate combination-invariant features via bidirectional knowledge transfer between complete and incomplete modal information. The transfer of knowledge from complete to missing enhances the performance of the incomplete modality combinations, while the reverse transfer reinforces the specificity of modality features. Furthermore, we randomly mask certain local features within re-

maining available modalities. The loss of local information within modalities encourages the model to either make inferences from the same modality itself or from other available modalities, or suffer a higher penalty for the absence of critical irreplaceable information. This encourages inter/intra-modal interaction and fosters the utilization of complementary specific information, thus prompting combination-specific representations. In addition, most existing incomplete multimodal models treat various missing modality combinations equally during training[10,21,26]. Nevertheless, distinct combinations of absent modalities hold varying degrees of information, resulting in diverse performance (imbalanced modality combinations). Treating them equally during training hinders the further optimization and improvement of weaker modality combinations. Therefore, we propose a curriculum learning strategy grounded in modality combinations to dynamically balance the training of imbalanced modality combinations, thus improving the model's representation ability for challenging combinations.

Our contributions can be summarized as follows: 1) We propose a novel and effective framework I2M2Net for incomplete multimodal skin disease classification, which utilizes inter/intra-modal feature masking self-distillation to learn unified combination-invariant features while prompting combination-specific representations. 2) We design a combination-based curriculum learning strategy to dynamically adjust the training of imbalanced modality combinations, promoting the robustness of the network for hard modality missing circumstances. 3) Experimental results on Derm7pt dataset [7] demonstrate the superiority of I2M2Net.

2 Method

The overall framework of proposed I2M2Net is presented in Fig. 1. It consists of three modality-specific encoders (denoted as E_c, E_d, E_m), a fusion block F for merging multimodal features (including a classification head for the final output prediction), and a mask generator M_g. Moreover, another branch F' in relation to F is introduced to perform self distillation. Note that the self-distillation and curriculum learning strategy are only utilized during the training phase and can be easily applied to other traditional multimodal models. And there is no additional structure and computational costs during inference compared to original multimodal networks. The details of each component are as follows.

2.1 Inter/Intra-modal Feature Masking Self-distillation

Knowledge distillation from a complete-modal network to the customized incomplete one has been proven effective in incomplete multimodal learning [4,6]. Inspired by the self-knowledge distillation strategy [8,1], which involves distilling knowledge within a model and utilizing it to train the model, we introduce an efficient self-distillation strategy which just introduces minimal parameters to implement knowledge distillation between complete and missing modality combinations within a unified model. It depends on inter-modal and intra-modal

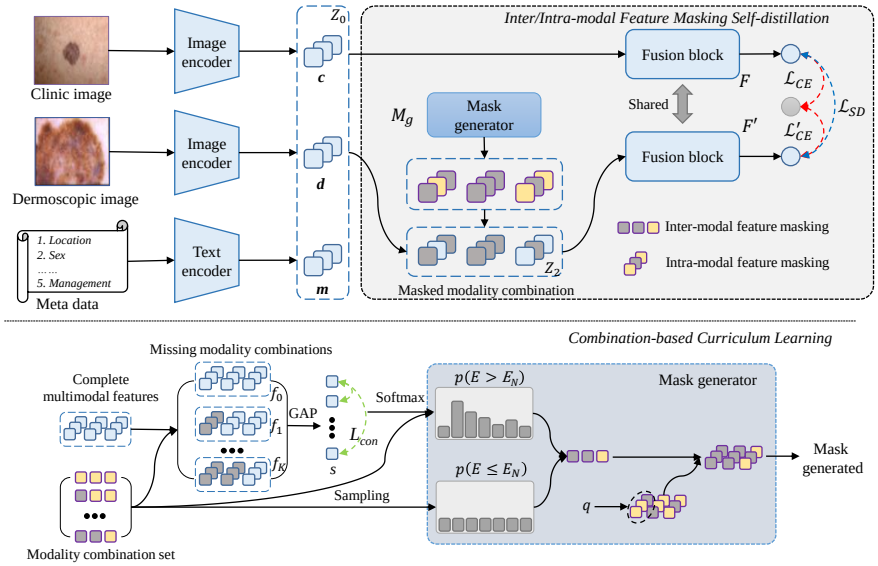

Fig. 1. The overview of the proposed framework I2M2Net. It mainly contains two important procedures, *i.e.* inter/intra-modal feature masking self-distillation and combination-based curriculum learning. F and F' share the same weights. E_N denotes the training epoch we set to start curriculum learning.

feature masking, which contributes to obtaining masked modality combinations. Self-distillation between F with complete modality information and F' with the masked modality features encourages combination-invariant/specific representations.

Concretely, F takes the original complete multimodal features (of the last stage in transformer-based modal encoders respectively) as input. Let $Z_0 = [c_1, \cdots, c_{n_c}, d_1, \cdots, d_{n_d}, m_1, \cdots, m_{n_m}] = [\mathbf{c}, \mathbf{d}, \mathbf{m}]$ be the concatenated multimodal features, where $c_i, d_i, m_i \in \mathbb{R}^C$ mean the features of clinical images, dermoscopic images, and metadata, respectively, n_c, n_d, n_m are the sequence length of each modality features, and C is the dimension (batch size is omitted for simplicity here). Another branch F' takes masked modality combinations Z_2 as input, which are obtained by employing two masking strategies concurrently on the complete Z_0.

Inter-modal feature masking. Random dropout of modality is utilized to simulate $2^J - 1$ different missing modality combinations (including the complete one), where J is the number of multiple modality and $J = 3$ in this paper. We first treat all combinations equally, and set the probability of each as $1/(2^J - 1)$. Let $M_{inter} = [\sigma_c, \sigma_d, \sigma_m]$ be the indicator of different missing combinations where $\sigma_i = 1$ represents the corresponding modality is existent and $\sigma_i = 0$ means not for $i \in c, d, m$. For example, if the clinical image modality frature is

missing, then $M_{inter} = [0, 1, 1]$, and the masked features can be defined as:

$$Z_1 = Z_0 \odot M_{inter} = [c'_1, \cdots, c'_{n_c}, d_1, \cdots, d_{n_d}, m_1, \cdots, m_{n_m}], \tag{1}$$

where c'_i means it is masked and $\odot$ is the shape-matched dot product operation.

Intra-modal feature masking. Meanwhile, for a given missing modality combination, we randomly masked local patches of the remaining at a rate of q to stimulate combination-specific representation learning. *i.e.*

$$Z_2 = Z_1 \odot M_{intra} = [c'_1, \cdots, c'_{n_c}, d_1, d''_2, \cdots, d_{n_d}, m''_1, m_2 \cdots, m''_{n_m}], \tag{2}$$

where m''_i, d''_i represent masked local features.

Self-distillation. By M_{inter}, the knowledge transfers between complete and incomplete modality combination, which promotes combination-invariant dependencies. This is bidirectional: the complete to missing modality combination enhances the performances of the incomplete ones, and that in the opposite direction potentially reinforces the specificity of modality features. On the other hand, the lack of local information after applying M_{intra} within a modality encourages the model to draw conclusions from the context of the modality or other available modalities, enhancing the interaction of information both between inter- and intra-modal. Otherwise, the absence of essential and unchangeable specific details results in a higher penalty, compelling the model to enhance its use of modal-specific information. Eventually, we apply the loss of self-distillation to force the class-wise consistency between F and F'

$$\mathcal{L}_{SD} = \left\| F(Z_0) - F'(Z_2) \right\|_1 . \tag{3}$$

The total loss of training is given as:

$$\mathcal{L} = \mathcal{L}_{CE}(y, F(Z_0)) + \mathcal{L}'_{CE}(y, F'(Z_2)) + \lambda \mathcal{L}_{SD}, \tag{4}$$

where y is the label, $\mathcal{L}_{CE}$ and $\mathcal{L}'_{CE}$ are cross entropy loss, and λ is the hyperparameter.

2.2 Combination-based Curriculum Learning

The information content of missing modalities is varying. While the absence of certain modalities can lead to a significant decline in performance, the absence of others may have only a minor effect (imbalanced modality combinations). Taking it into consideration, we design a curriculum learning strategy to dynamically harmonize the training of various strong or weak masked modality combinations. The training is divided into two stages. (a) $E \leq E_N$ when our model is trained normally, and (b) $E > E_N$ when we assign different training weights to imbalanced masked modality combinations according to their degrees of capabilities. Following the normal framework of difficulty measurer and training scheduler

in curriculum learning [18], a difficulty measurer to evaluate the competency level across various modality combinations is designed. Intuitively, the complete modalities, encompassing the most information within various modality combinations, is perceived to possess the highest capacity. Therefore, we utilize the consistency between complete modalities and other missing combinations as an indicator of capability, where higher consistency signifies enhanced capability. We can obtain them by

$$s_i = L_{con}(GAP(f_0), GAP(f_i)), i = 1, 2, \ldots, K, \tag{5}$$

where f_0 is the complete modalities, K and f_i represents the num of missing modality combinations and the i-th one respectively. Here, $K = 2^J - 2$ and it is equal to 6 for three modality inputs in this paper. GAP is the global average pooling operation, and the cosine similarity is chose as L_{con}.

For training scheduler, in Sec. 2.1, we implemented self-distillation by training with randomly chosen various modality combinations as branch inputs. Here, we adjust the probabilities p_i of different missing modality combinations appearing in training based on s_i. As s_i increases, p_i decreases. *i.e.*

$$p_i = Softmax(1 - s_i), i = 0, 1, \cdots, K, s_0 = Max(s_i), \tag{6}$$

where $Softmax$ is the Softmax function, and Max denotes acquiring the largest value from a set $\{s_i\}_{i=1}^{K}$. Let s_0 represents the consistency of complete modalities itself, set to $Max(s_i)$.

2.3 Modality Deficiency-aware Fusion

Transformers have demonstrated significant advantages in multimodal fusion tasks [16,5]. However, because the self-attention operation is highly sensitive to missing data, the lack of modalities may result in a noticeable decrease in model performance. Inspired by [13], we adopt stacked transformer blocks in F with masked multi-head self-attention to adaptively fuse available modal information. It avoids the impact of missing modality on self-attention, and thereby we call it modality deficiency-aware fusion.

3 Experiments

3.1 Datasets and Evaluation Metrics

We evaluate the I2M2Net on the publicly available multimodal skin lesion dataset Derm7pt [7], which consists of 1011 cases in total with three modalities: dermoscopic image (der), clinical image (cli) and patient meta-data (meta). These 1011 cases have been officially divided into three sets: 413 cases for training, 203 cases for validation, and 395 cases for testing purposes. Larger dataset will be considered in future work. Accuracy (Acc) and F1-score (shown in supplementary material) are adopted as the evaluation metrics for evaluation.

Table 1. Comparison results on Derm7pt datasets (Acc%). # denotes complete multimodal method. ∘ and • represent missing and available modality, respectively. Avg. reports the average value of 7 modality combinations. Data format: avg (std)

der	•	∘	∘	•	•	∘	•	Avg.
cli	∘	•	∘	•	∘	•	•	
meta	∘	∘	•	∘	•	•	•	
Concat #	66.58 (4.47)	52.56 (6.46)	65.82 (6.46)	67.24 (5.64)	76.86 (1.95)	72.15 (1.68)	76.45 (1.42)	69.07 (2.43)
Remixformer[19] #	63.04 (4.88)	55.95 (1.66)	57.62 (3.82)	69.16 (2.62)	77.27 (2.20)	67.24 (2.00)	79.54 (0.90)	67.12 (2.10)
Tformer[23] #	65.06 (5.14)	55.19 (1.05)	58.48 (6.14)	70.78 (3.01)	77.42 (2.78)	65.17 (6.62)	78.84 (1.85)	68.23 (0.92)
MMIN[26]	65.57 (2.25)	59.09 (3.33)	70.99 (1.10)	67.49 (3.28)	76.66 (0.79)	74.89 (0.94)	76.00 (1.18)	70.60 (1.14)
MP[10]	64.91 (1.77)	60.30 (1.94)	71.29 (1.29)	67.79 (1.10)	75.34 (1.55)	**75.85 (1.35)**	76.20 (0.77)	70.74 (1.08)
LCKD[17]	70.18 (1.59)	66.88 (0.94)	71.14 (0.70)	71.09 (1.39)	76.86 (0.88)	75.14 (0.81)	78.23 (0.36)	72.79 (0.48)
TATE[21]	70.28 (0.65)	64.96 (1.13)	71.14 (0.42)	71.24 (1.14)	**77.92 (1.22)**	74.78 (1.21)	78.33 (0.99)	73.16 (0.43)
I2M2Net	**71.60 (2.54)**	**68.56 (1.08)**	**71.65 (1.95)**	**73.32 (1.24)**	77.87 (0.63)	75.49 (0.98)	**79.64 (0.97)**	**74.02 (0.93)**

3.2 Implementation Details

The proposed method is implemented on an NVIDIA A100 GPU with PyTorch. Our backbone for image data is a regular Swin-T/224 [12], and for text data is the typical Transformer encoder[16]. The data augmentation includes random flip, rotation, clipping, dwp [19].We use Adam [9] as the optimizer with the weight decay 1e-4 and the initial learning rate of 0.0001 for 100 epochs. The number and dimension of the transformer blocks in Fusion module are 3 and 768. The hyperparameters q, E_N, λ are set to 0.2, 20, and 1 respectively. The model achieving the highest Acc on validation set with a 30% missing rate is selected for testing. Note that different training using the same missing dataset. Additionally, we perform experiments in 5 independent runs and present the average value and standard deviation (avg/std) in the results.

3.3 Comparison Results

To evaluate the proposed I2M2Net, we compare our method with 7 advanced methods on Derm7pt dataset. Among them, concat, Remixformer [19] and Tformer [23] are complete multimodal classification method, while MMIN [26], MP [10], LCKD [17] and TATE [21] are state-of-the-art approaches for incomplete multimodal learning. For fair comparison, we re-train all methods based on public code repositories in the same experimental environment. The results of all seven modality combinations are shown in Table 1. It can be seen that our approach significantly outperforms the other comparative methods across the majority

Table 2. Ablation study on Derm7pt datasets (Acc%). Avg. reports the average value of 7 modality combinations. Data format: avg (std)

der cli meta	● ○ ○	○ ● ○	○ ○ ●	● ● ○	● ○ ●	○ ● ●	● ● ●	Avg.
Baseline	69.31 (1.15)	62.68 (1.86)	68.91 (2.28)	69.12 (1.05)	76.71 (1.10)	74.53 (1.45)	77.87 (1.09)	71.31 (0.88)
+Inter-Mask	66.38 (2.74)	65.52 (1.75)	72.50 (0.94)	70.43 (0.52)	75.75 (0.90)	74.43 (0.89)	77.17 (0.54)	71.74 (0.40)
+Intra-Mask($q = 0.2$)	71.60 (0.96)	65.77 (1.85)	65.57 (6.16)	72.30 (1.14)	76.10 (1.27)	73.31 (2.58)	77.92 (1.98)	71.80 (1.52)
+Inter&Intra-Mask($q = 0.2$)	71.45 (2.27)	65.32 (1.98)	71.95 (1.27)	72.20 (2.62)	76.00 (1.60)	75.19 (0.53)	78.99 (0.98)	73.01 (0.78)
+Inter&Intra-Mask($q = 0.1$)+CCL	**72.91** **(1.95)**	67.50 (1.10)	71.95 (0.85)	72.51 (1.61)	77.62 (1.43)	75.09 (1.40)	78.73 (0.99)	73.76 (0.43)
+Inter&Intra-Mask($q = 0.2$)+CCL	71.60 (2.54)	**68.56** **(1.08)**	71.65 (1.95)	**73.32** **(1.24)**	**77.87** **(0.63)**	**75.49** **(0.98)**	**79.64** **(0.97)**	**74.02** **(0.93)**
+Inter&Intra-Mask($q = 0.3$)+CCL	71.65 (1.87)	68.05 (0.74)	71.90 (1.56)	72.81 (1.01)	76.81 (1.07)	75.09 (1.67)	78.73 (1.05)	73.58 (0.65)
+Inter&Intra-Mask($q = 0.5$)+CCL	70.99 (0.57)	66.02 (2.88)	**72.91** **(1.05)**	72.20 (1.64)	77.01 (1.06)	73.97 (2.23)	78.43 (1.45)	73.08 (1.13)

of modal combinations(*i.e.* 5 out of 7 and Avg.). For instance, we improve the average Acc scores by 1.35% compared with TATE (p=0.025<0.05) and 1.23% compared with LCKD (p=0.003<0.05). Note worthy that our approach demonstrates a great gain 3.6% and 1.68% in the weakest performing modality combination(*i.e.* [der*, cli, meta*], where * represents missing) compared to Tate and LCKD, respectively. This illustrates that our combination-based curriculum learning has indeed effectively enhanced the performance of weak modal combinations. Meanwhile, in comparison to three fully modal methods, our approach still shows good enhancement (*i.e.* 3.19%, 0.1% and 0.8%, respectively) in performance when employing full multimodal data input, indicating the effectiveness of I2M2Net even for complete multimodal skin lesion classification.

3.4 Ablation Study

To verify the effectiveness of each component in I2M2Net, we further conduct detailed ablation experiments, the results of which are reported in Table 2. Initially, the original single-branch network with the modality deficiency-aware fusion module is adopted as the "Baseline". We can observed that simply incorporating the inter-modal or intra-modal feature masking self-distillation alone acquires not satisfactory performance, falling short of both Tate and LCKD. The combination of these two elements enhances the combination-invariant and combination-specific representations, thereby enhancing the overall performance of classification. Furthermore, the introduction of combination-based curriculum learning strategy has remarkably improved the model's robustness to weak modality combinations. In Table 2, the Acc of 6th experiment outperforms by 3.24% in the weakest modality combinations, [der*, cli, meta*]. In addition, we investigate the impact of the hyperparameter q (*i.e.* the ratio of intra-mask). It

can be seen that the model performs well in most modality combinations when q being set as 0.2. We report that a higher degree of masking complicates self-distillation training, whereas a lower level of masking does not effectively enhance combination-specific representations, resulting in a decrease in performance.

4 Conclusion

In this study, a novel and effective framework named I2M2Net is presented for incomplete multimodal learning. We first incorporate inter-modal and intra-modal feature masking self-distillation techniques to enhance combination-specific representations and promote learning of combination-invariant dependencies, thus improving overall robustness to incomplete multimodal data. Furthermore, we design a combination-based curriculum learning algorithm to assist the network in adapting to weak and challenging modality combinations. Extensive experiments conducted on the Derm7pt dataset demonstrate the effectiveness and superiority of I2M2Net compared to other contemporary approaches.

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

References

1. Baevski, A., Hsu, W.N., Xu, Q., Babu, A., Gu, J., Auli, M.: Data2vec: A general framework for self-supervised learning in speech, vision and language. In: International Conference on Machine Learning. pp. 1298–1312. PMLR (2022)
2. Balch, C.M., Gershenwald, J.E., Soong, S.j., Thompson, J.F., Atkins, M.B., Byrd, D.R., Buzaid, A.C., Cochran, A.J., Coit, D.G., Ding, S., et al.: Final version of 2009 ajcc melanoma staging and classification. Journal of clinical oncology **27**(36), 6199 (2009)
3. Cai, L., Wang, Z., Gao, H., Shen, D., Ji, S.: Deep adversarial learning for multi-modality missing data completion. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 1158–1166 (2018)
4. Chen, C., Dou, Q., Jin, Y., Liu, Q., Heng, P.A.: Learning with privileged multimodal knowledge for unimodal segmentation. IEEE transactions on medical imaging **41**(3), 621–632 (2021)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
6. Hu, M., Maillard, M., Zhang, Y., Ciceri, T., La Barbera, G., Bloch, I., Gori, P.: Knowledge distillation from multi-modal to mono-modal segmentation networks. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23. pp. 772–781. Springer (2020)
7. 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)

8. Kim, K., Ji, B., Yoon, D., Hwang, S.: Self-knowledge distillation: A simple way for better generalization. arXiv preprint arXiv:2006.12000 **3**, 1 (2020)
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
10. Lee, Y.L., Tsai, Y.H., Chiu, W.C., Lee, C.Y.: Multimodal prompting with missing modalities for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14943–14952 (2023)
11. Liu, J., Liu, X., Zhang, Y., Zhang, P., Tu, W., Wang, S., Zhou, S., Liang, W., Wang, S., Yang, Y.: Self-representation subspace clustering for incomplete multi-view data. In: Proceedings of the 29th ACM international conference on multimedia. pp. 2726–2734 (2021)
12. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
13. Recasens, A., Lin, J., Carreira, J., Jaegle, D., Wang, L., Alayrac, J.b., Luc, P., Miech, A., Smaira, L., Hemsley, R., et al.: Zorro: the masked multimodal transformer. arXiv preprint arXiv:2301.09595 (2023)
14. Stroud, J., Ross, D., Sun, C., Deng, J., Sukthankar, R.: D3d: Distilled 3d networks for video action recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 625–634 (2020)
15. Tran, L., Liu, X., Zhou, J., Jin, R.: Missing modalities imputation via cascaded residual autoencoder. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1405–1414 (2017)
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
17. Wang, H., Ma, C., Zhang, J., Zhang, Y., Avery, J., Hull, L., Carneiro, G.: Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 216–226. Springer (2023)
18. Wang, X., Chen, Y., Zhu, W.: A survey on curriculum learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(9), 4555–4576 (2021)
19. Xu, J., Gao, Y., Liu, W., Huang, K., Zhao, S., Lu, L., Wang, X., Hua, X.S., Wang, Y., Chen, X.: Remixformer: A transformer model for precision skin tumor differential diagnosis via multi-modal imaging and non-imaging data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 624–633. Springer (2022)
20. Yang, Y., Xie, F., Zhang, H., Wang, J., Liu, J., Zhang, Y., Ding, H.: Skin lesion classification based on two-modal images using a multi-scale fully-shared fusion network. Computer Methods and Programs in Biomedicine **229**, 107315 (2023)
21. Zeng, J., Liu, T., Zhou, J.: Tag-assisted multimodal sentiment analysis under uncertain missing modalities. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1545–1554 (2022)
22. Zhang, Y., Chen, J., Wang, K., Xie, F.: Ecl: Class-enhancement contrastive learning for long-tailed skin lesion classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 244–254. Springer (2023)

23. Zhang, Y., Xie, F., Chen, J.: Tformer: A throughout fusion transformer for multimodal skin lesion diagnosis. Computers in Biology and Medicine **157**, 106712 (2023)
24. Zhang, Y., Xie, F., Song, X., Zheng, Y., Liu, J., Wang, J.: Dermoscopic image retrieval based on rotation-invariance deep hashing. Medical Image Analysis **77**, 102301 (2022)
25. Zhang, Y., Xie, F., Song, X., Zhou, H., Yang, Y., Zhang, H., Liu, J.: A rotation meanout network with invariance for dermoscopy image classification and retrieval. Computers in Biology and Medicine **151**, 106272 (2022)
26. Zhao, J., Li, R., Jin, Q.: Missing modality imagination network for emotion recognition with uncertain missing modalities. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 2608–2618 (2021)
27. Zhou, T., Ruan, S., Hu, H.: A literature survey of mr-based brain tumor segmentation with missing modalities. Computerized Medical Imaging and Graphics **104**, 102167 (2023)