2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-3748-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/BIBM58861.2023.10385898

Multi-modal Contrastive-Generative Pre-training for Fine-grained Skin Disease Diagnosis

Liangdi Ma[1, 2], Jun Zhao[3], Guoxin Wang[3], Yuchen Guo[2], and Feng Xu[1, 2*]
[1]School of Software, [2]BNRist, Tsinghua University, Beijing, China
[3]JD Health International Inc., Beijing, China
mld21@mails.tsinghua.edu.cn, {zhaojun10, wangguoxin14}@jd.com, {yuchen.w.guo, xufeng2003}@gmail.com

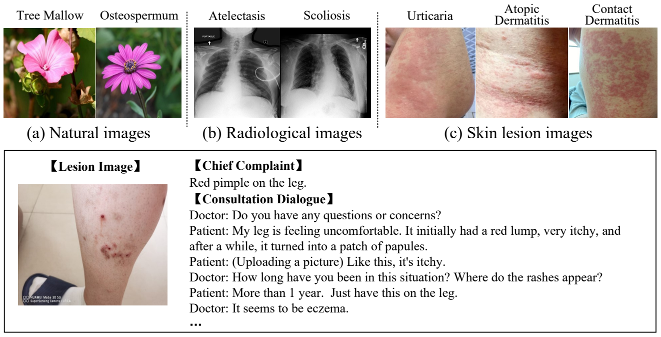
***Abstract*—Vision-language pre-training (VLP) leverages easily accessible image-text pairs instead of high-cost expert-annotated labels for pre-training, which has achieved promising performance and attracted considerable attention. There are many works on *coarse-grained* VLP in natural and medical radiology applications. However, as a common problem, the *fine-grained* setting is still unexplored, especially in medical applications like skin disease diagnosis. In fine-grained cases, the visual appearance of different objects is highly similar, and the language information may be sparse and noisy, both remarkably increasing the difficulty of learning effective features by VLP. In this paper, we address these difficulties and propose a novel Multi-level Multi-modal Contrastive-Generative (M2CG) pre-training method. M2CG has 1) a feature-level multi-modal contrastive module to learn fine-grained features via semantic knowledge, and 2) a semantic-level cross-modal generation module to enforce the model to capture key and discriminative features. We construct a multi-modal skin disease dataset, containing user-taken lesion photos, chief complaints, and consultation dialogues, to perform VLP with M2CG and evaluate the performance on three public skin disease benchmarks and a fine-grained dataset with 64 categories collected from real-world applications. M2CG outperforms the state-of-the-art VLP methods by up to 11.11% in diagnosis accuracy, yielding consistent and significant promotion and facilitating skin disease diagnosis. To our knowledge, this is the first VLP study presented for fine-grained skin disease diagnosis. We believe that the success of M2CG will inspire more innovations in fine-grained VLP for medical practice.**

***Index Terms*—Skin disease diagnosis, Multi-modal pre-training, Vision-language pre-training**

I. Introduction

As the first barrier of the human body, skin is a vital part of the human immune system [1]. An immediate and accurate diagnosis from lesion images is essential for further treatment and prognosis of patients. The advancement of deep learning (DL) achieves impressive performance in medical applications [2], [3] and presents promising opportunities for skin disease diagnosis, which has been an important research topic and attracted considerable attention recently [4]–[6].

A major challenge for DL is the collection of enormous annotated data, which is especially difficult in medical applications. As it is cheap and flexible to collect image-text pairs (say, by Web data [7]), vision-language pre-training (VLP) shows superiority to offset the over-reliance on annotations of deep learning, which gains wide attention to facilitate subsequent tasks [7], [8]. Existing studies suggest the visual and linguistic features learned from large-scale image-text pairs benefit both vision and vision-language tasks [9], [10]. A few latest works introduce VLP to medicine with inherently multi-modal data in practice(e.g., radiology reports) and achieve promising results on various radiology-related tasks [11]–[13].



(a) Natural images (b) Radiological images (c) Skin lesion images

【Chief Complaint】
Red pimple on the leg.
【Consultation Dialogue】
Doctor: Do you have any questions or concerns?
Patient: My leg is feeling uncomfortable. It initially had a red lump, very itchy, and after a while, it turned into a patch of papules.
Patient: (Uploading a picture) Like this, it's itchy.
Doctor: How long have you been in this situation? Where do the rashes appear?
Patient: More than 1 year. Just have this on the leg.
Doctor: It seems to be eczema.
...

(d) Example of (image, dialogue, chief complaint) triple of skin disease

Fig. 1. (a-c). Examples of natural, radiological, and fine-grained skin lesion images. (d). Example of the lesion image, chief complaint, and doctor-patient dialogue in the electronic record for pre-training.

However, pre-training for skin disease diagnosis is much more challenging. Firstly, the skin images are intrinsically *fine-grained*, different from natural images or radiology images which are coarse-grained. As Fig. 1 shows, the skin lesions from different diseases share highly similar visual appearances, making it more necessary for pre-training to identify the discriminative features from the skin lesion images. Secondly, unlike the textual captions grabbed from radiology reports, collecting dense and accurate captions for skin lesion images is laborious. A feasible way to acquire lesion-related text is from the chief complaints given by the patients and the consultation dialogues between doctors and patients. However, these texts have sparse and inaccurate semantics in the contents (Fig. 1d). Due to the challenges in both vision and language perspectives, it is difficult and sub-optimal for existing VLP approaches to learn the fine-grained features for skin disease diagnosis.

To achieve a better understanding of fine-grained skin lesion images, in this paper, we collect a large-scale multi-modal skin disease dataset and propose a novel Multi-level Multi-modal Contrastive-Generative (M2CG) pre-training framework for fine-grained skin disease diagnosis. M2CG has two key novel components: 1) multi-modal contrastive, which exploits discriminative medical insights introduced from consultation dialogues to promote the understanding of the fine-grained lesion

*Corresponding author

979-8-3503-3748-8/23/$31.00 ©2023 IEEE

images, and 2) cross-modal chief complaint generation, which provides explicit semantic-level supervision to further encourage the awareness of key features in the lesion images and the complicated dialogue texts. Besides, we also use single-modal contrastive to learn the generic visual features and improve the robustness to different image capture conditions (illumination, view, etc). We perform pre-training on the large-scale multi-modal skin disease dataset and conduct extensive experiments on three public skin disease diagnosis benchmarks [14]–[16] and an in-house 64-class dataset. M2CG consistently yields superior results to the state-of-the-art (SOTA) methods even in the extremely fine-grained scenario, suggesting that M2CG is beneficial to fine-grained feature extraction and promising for dermatological applications. The main contributions of our paper are summarized as follows:

1. We propose a multi-modal pre-training framework M2CG, exploiting intra- and inter-modal knowledge simultaneously by feature-level contrastive and semantic-level generative learning to enhance the fine-grained features capture.
2. We construct a large-scale multi-modal pretraining dataset containing skin lesion images, patient's chief complaints, and doctor-patient consultation dialogue, and a multi-class skin disease diagnosis dataset to assess the performance.
3. M2CG achieves state-of-the-art performance on three public datasets and yields promising performance in real-world applications, suggesting that M2CG can provide comprehensive representations for fine-grained features.

II. Method

In this paper, we propose a VLP approach M2CG consisting of four key modules, i.e. Single-Modal Contrastive (SMC), Multi-Modal Contrastive (MMC), Multi-Modal Fusion (MMF), and Cross-Modal Generation (CMG), to extract features from the skin lesion images, the patient's chief complaints, and the patient-doctor consultation dialogues to facilitate skin disease diagnosis. M2CG first acquire two augmented views of each lesion image, tokenize the consultation dialogue and chief complaint, and extract multi-modal features from lesion images and consultations individually. We perform SMC and MMC for *feature-level* alignment, maximizing the consistency between paired features to learn transform-invariant representations and introduces semantic knowledge. Besides, we integrate CMG to perform multi-task pre-training, which provides *semantic-level* supervision to enhance the awareness of discriminative features by explicitly predicting the chief complaints based on the multi-modal features. The overall pre-training framework is shown in Fig. 2,

A. Pre-processing

For vision input, during pre-training, given a lesion image $I_i \in \mathcal{R}^{3\times H\times W}$, we first resize the image to 512×512 and perform a set of random transformations including crop, horizontal flip, rotation, horizontal and vertical translation, scale, and Gaussian blurring twice to create two augmented views (I_{i1}, I_{i2}). The transformed images are further resized to 224×224 resulting the input $I_{ik} \in \mathcal{R}^{3\times224\times224}$ of the network. During finetuning , we employ the same augmentations as pre-training to enhance the robustness of M2CG. During assessment, we directly resized the input image to 224×224 without augmentation for accurate evaluation.

For the text input including consultation dialogues and chief complaints, we tokenize the texts into a sequence of tokens and set the maximum length to 512 for the consultation dialogues and 60 for the chief complaints, respectively. The over-long sequences are truncated and the over-shorts are padded with certain tokens to form the input of the networks.

B. Multi-modal Feature Extraction

To extract and leverage the high-level semantic information in the multi-modal data, M2CG utilizes a CNN-based vision encoder E_v and an attention-based text encoder E_t to encode the input multi-modal data.

Given the augmented views I_{i1} and I_{i2}, M2CG employs vision encoder E_v to encode each image into a sequence of patch embeddings V_{ik}. Specifically, E_v maps the input view I_{ik} to a feature map $F \in \mathcal{R}^{H_{\text{output}}\times W_{\text{output}}\times d_v}$. The output feature map is further reshaped to acquire the patch embeddings $V_{ik} \in \mathcal{R}^{H_{\text{output}}W_{\text{output}}\times d_v}$, where each d_v-dimensional feature represents a patch in I_{ik}. Subsequently, the global feature $v_{ik} \in \mathcal{R}^{d_v}$ of I_{ik} can be acquired by performing global average pooling to the patch embeddings V_{ik} as:

$$v_{ik} = \text{GAP}(V_{ik}), \text{where } V_{ik} = E_v(I_{ik}) \tag{1}$$

where GAP refers to the global average pooling operation, and k denotes the k-th augmented view of the lesion image.

For text encoding, we build the text encoder E_t using the Transformer encoder layer [17] to extract the text representation. Following the practice of natural language processing, given L_t input tokens, two embedding layers are applied to the input tokens to extract the token and positional embeddings. The token and positional embeddings are added together to inject position information of the dialogue text T_i into the input token embeddings $U_i \in \mathcal{R}^{L_t\times d_t}$. U_i is fed into the text encoder E_t consisting of multiple stacked multi-head self-attention layers to compute the correlation weights between tokens [17]. The multi-head attention can be formulated as:

$$\begin{aligned} U_i' &= \text{Concat}(attn_1, ..., attn_H)W^O \\ \text{where} \quad attn_h &= \text{softmax}(\frac{QW_h^Q(KW_h^K)^\top}{\sqrt{d_t/H}})VW_h^V \end{aligned} \tag{2}$$

where Q, K, V refer to the queries, keys, and values, set as the input token embeddings, and W^Q, W^K, W^V denote the corresponding learnable parameter matrices, to perform the self-attention computation. Concat refers to the concatenating operation. We utilize $H = 8$ parallel attention heads in E_t, where the output of each head $attn_h$ are concatenated and projected by the parameter matrix W^O to obtain the final output embeddings $U_i' \in \mathcal{R}^{L_t\times d_t}$ of each multi-head attention layer. The output embeddings of each multi-head attention layer will be further input to a fully connected feed-forward network FFN, which consists of two linear layers with ReLU

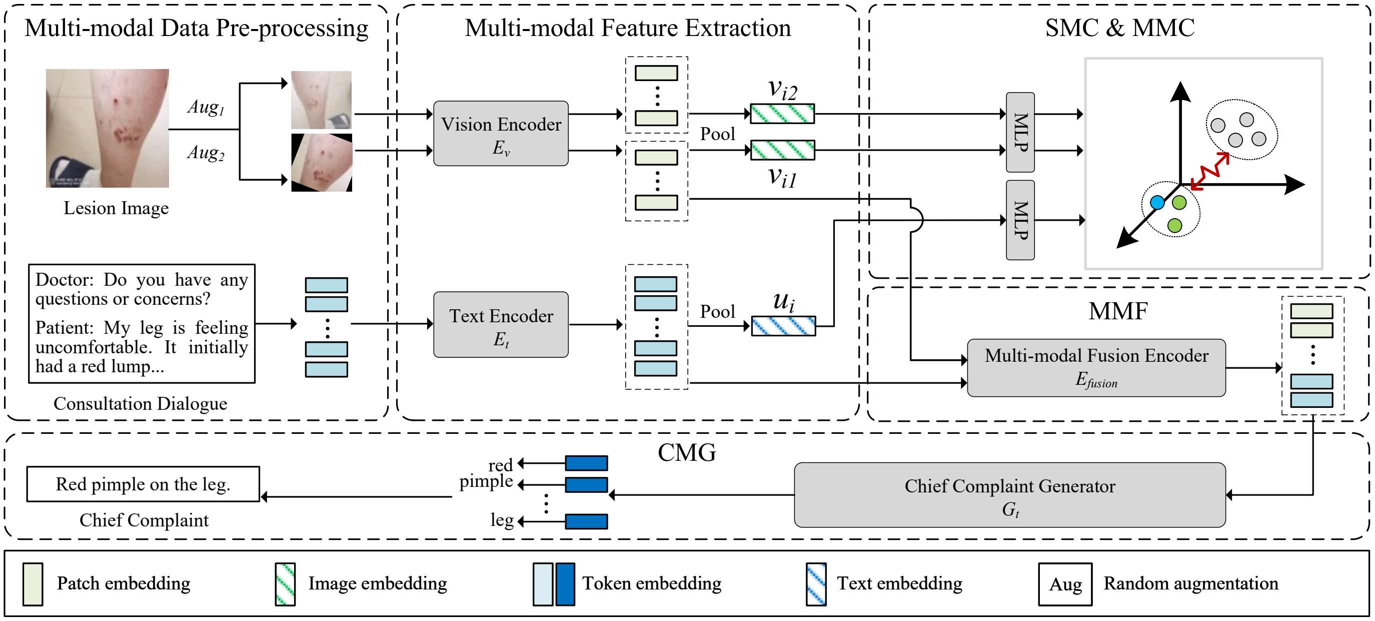


Fig. 2. Framework of M2CG. The dotted boxes outline the key modules of pre-training. SMC: Single-Modal Contrastive. MMC: Multi-Modal Contrastive. MMF: Multi-Modal Fusion. CMG: Cross-Modal Generation(CMG).

activation. Finally, the global feature u_i of the whole input dialogue is extracted by mask-based average pooling to the output token embeddings of E_t with the attention masks as:

$$u_i = \mathrm{MAP}(\mathrm{FFN}(U_i') \odot M_i) \tag{3}$$

where M_i is the attention mask specifying attended tokens. MAP refers to the mask-based average pooling operation, which is performed on the attended tokens only to integrate the token embeddings into the final dialogue features.

C. Single-Modal Contrastive

Unlike regular radiological examinations, we observed that skin lesion images are usually photoed under uncontrollable surroundings and conditions. Inspired by the transformation-invariant benefits of self-supervised learning methods [18], we utilize single-modal contrastive(SMC) to improve the robustness of the image encoder to the lesion images. SMC takes advantage of the inherent distribution of images to extract generic visual representations via self-supervised learning.

Specifically, given the encoded features $(v_{i1}, v_{i2}) \in \mathcal{R}$ of the paired views, we utilize an additional non-linear projection head MLP to map v_{ik} to a modal-shared latent space to extract the visual embeddings $\hat{v}_i$. The objective of SMC is to maximize consistency between the visual embeddings of the paired views within the mini-batch:

$$\mathcal{L}_{\mathrm{SMC}} = -\frac{1}{2N}\sum_{i}^{N}\sum_{k}^{2}\log\frac{\exp(\mathrm{sim}(\hat{v}_{i1}, \hat{v}_{i2})/\tau_s)}{\sum_{j\neq i}\exp(\mathrm{sim}(\hat{v}_{ik}, \hat{v}_j)/\tau_s)} \tag{4}$$

where N and τ_s refer to the batch size and the temperature parameter (set to 0.1 in this paper), respectively. $\hat{v}_j$ represents the other samples in the mini-batch. sim denotes the *cosine* similarity measurement, which can be formulated as:

$$\mathrm{sim}(u, v) = u \cdot v/(\|u\|\|v\|) \tag{5}$$

SMC yields the visual features $\hat{v}_i$ from skin lesion images as the fundamental block of the following multi-modal analysis.

D. Multi-Modal Contrastive

To leverage the domain knowledge of skin disease, M2CG explores medical consultations, which contain rich semantics and detailed descriptions of the paired image. Feature-level alignment is proven to be an effective technique for collectively acquiring multi-modal knowledge [7]. Therefore, M2CG employs the consultation dialogues via multi-modal contrastive(MMC) to introduce medical insights to the paired lesion images and facilitate fine-grained feature extraction.

Given the global features of the dialogue, an additional MLP head is applied to project u_i into the modal-shared latent space to extract the textual embeddings $\hat{u}_i$. With the visual embeddings $\hat{v}_i$ and the textual embeddings $\hat{u}_i$, the target of MMC is to align and maximize the agreement between the embeddings of corresponding (image, dialogue) pairs to preserve mutual information in each mini-batch. We compute the asymmetric contrastive loss to the image and text as [13]:

$$\begin{aligned}\mathcal{L}_{\mathrm{MMC}} = -\frac{1}{2N}\sum_{i}^{N}\sum_{k}^{2}(\lambda\log\frac{\exp(\mathrm{sim}(\hat{v}_{ik}, \hat{u}_i)/\tau_m)}{\sum_j\exp(\mathrm{sim}(\hat{v}_{ik}, \hat{u}_j)/\tau_m)} \\ + (1-\lambda)\log\frac{\exp(\mathrm{sim}(\hat{v}_{ik}, \hat{u}_i)/\tau_m)}{\sum_j\exp(\mathrm{sim}(\hat{u}_i, \hat{v}_{jk})/\tau_m)})\end{aligned} \tag{6}$$

where τ_m and λ denote the multi-modal temperature parameter and the asymmetric multi-modal loss weight, which are set to 0.1 and 0.75, respectively. Similarly, $\hat{v}_j$ and $\hat{u}_j$ denote the rest of visual or textual embeddings within a mini-batch. MMC introduces detailed semantics as auxiliary guidance for M2CG to achieve better fine-grained visual feature learning.

E. Multi-Modal Fusion

To capture the local relationship between intra- and inter-modal representations, M2CG further employs a multi-modal fusion encoder (MMF) to promote a more fine-grained understanding of lesion image and dialogue.

Overall, the multi-modal embeddings can be expressed as:

$$X_{\mathrm{mm}} = \mathrm{Concat}(V_{i1}, U_i) \tag{7}$$

where V_{i1} and U_i denote the output sequence of patch- and token-level embeddings from the vision and text encoder. The multi-modal embeddings are fed into a multi-modal fusion encoder E_{fusion} performing multi-head self-attention to model the local interactions between intra- and cross-modal features. The calculation mechanism of MMF can be formulated as aforementioned (2), while the input values of Q, K, V are set as the concatenated multi-modal embeddings exceptionally. The produced multi-modal features of MMF will be further fed into the following CMG module as references to summarize the key content in the image and dialogue.

F. Cross-Modal Generation

Unlike regular image captions or radiology reports, the content of medical consultation is presented as more complicated, due to its indirect, sparse, and inaccurate presentation. This challenging feature makes aligning multi-modal representations in a latent space insufficient and sub-optimal for fine-grained VLP. To encourage the awareness of the essential and discriminative features in both lesion images and consultation dialogues, we incorporate a novel cross-modal chief complaint generation(CMG) module, which provides semantic-level supervision explicitly in a generative manner to promote fine-grained feature learning.

The chief complaint generator G_t is built upon Transformer decoder [17] to predict the chief complaint. Each layer of G_t orderly performs masked-self-attention and cross-attention computation followed by a FFN to extract the semantic information from the multi-modal embeddings. Specifically, the input tokens of the chief complaint C_i are embedded and fed into G_t. G_t shares a similar structure to E_t with two distinctions. 1) To prevent information leakage from the unseen tokens, the input token embeddings are first masked out to preserve the auto-regressive property. Thus, the self-attention (2) is performed to the masked token embeddings in G_t. 2) An additional multi-head cross-attention layer, which computes multi-head attention like (2) but takes the encoded multi-modal embeddings X_{mm} from MMF as keys and values and the token embeddings S_i' as queries, is inserted before the FFN. Therefore, the cross-attention computation on (2) can be modified and reformulated as:

$$attn_h = \mathrm{softmax}(\frac{S_i' W_h^Q (X_{mm} W_h^K)^\top}{\sqrt{d_{mm}/H}}) X_{mm} W_h^V \tag{8}$$

where d_{mm} is the dimension of $X_m m$. The cross-attention layer allows G_t to aggregate and decode the informative features from the encoded multi-modal embeddings, utilizing information about lesion images and consultation dialogues as guidance to generate the chief complaints.

The target of CMG is to generate the chief complaint C_i based on the features extracted from the corresponding lesion image I_i and consultations T_i. Therefore, given a triple of (I_i, T_i, C_i), the objective of CMG is to maximize the conditional log-likelihood probability of the predicted $\hat{C}_i$ in an auto-regressive manner:

$$\mathcal{L}_{\mathrm{CMG}} = -\frac{1}{N} \sum_{i}^{N} \sum_{l}^{L_c} \mathrm{logP}(c_l | c_1, ..., c_{l-1}, I_i, T_i) \tag{9}$$

where L_c is the length of C_i. $\mathcal{L}_{\mathrm{CMG}}$ provides strong supervision to capture the discriminative features in the lesion images and dialogues for fine-grained skin disease.

Accordingly, the overall objective of M2CG is:

$$\mathcal{L}_{\mathrm{M2CG}} = \lambda_{\mathrm{SMC}} \mathcal{L}_{\mathrm{SMC}} + \lambda_{\mathrm{MMC}} \mathcal{L}_{\mathrm{MMC}} + \lambda_{\mathrm{CMG}} \mathcal{L}_{\mathrm{CMG}} \tag{10}$$

where $\lambda_{\mathrm{SMC}}, \lambda_{\mathrm{MMC}}, \lambda_{\mathrm{CMG}}$ denote the weights of SMC, MMC, and CMG and empirically set to 2.0, 1.0, and 1.0.

III. Experiments

A. Dataset and Implementation Details

Pre-training Dataset. Considering the domain shift and fine-grained problems, existing multi-modal datasets concentrating on the natural or radiology data are sub-optimal for skin disease diagnosis. Thus, we construct a large-scale multi-modal skin disease dataset sourced from the electronic records from the Internet dermatology specialist hospital for pre-training. We first collect about 796k medical records and filter out those with skin-irrelevant images, overlong texts, or incomplete information. Finally, we obtain a pre-training dataset containing 500k skin-related medical records, each of which consists of a patient-uploaded lesion image, a chief complaint, and a doctor-patient consultation dialogue, as Fig. 1d shows.

Downstream Dataset. To verify the effectiveness of M2CG, we conduct experiments on three public skin disease classification benchmarks and an extremely fine-grained dataset Skin64 collected from real applications. **MSLD** [15]: a binary monkeypox skin lesion detection dataset with 102 images belonging to "Monkeypox" and 126 for "Others". **MSID** [16]: a large-scale web-scrapped skin image database containing 5 different skin diseases (Monkeypox, Chickenpox, Smallpox, Cowpox, and Measles) and healthy skin images. We employ MSLD and MSID sourced from *Kaggle* with 3,152 and 39,396 augmented images for 5-fold cross-validation, respectively. **DermNet23** [14]: a multi-class dataset containing 19,559 images of 23 skin disease types. We follow the official split and sample 10% training data for validation. **Skin64**: a multi-class dataset with 10,533 lesion images and 64 classes of skin disease. Skin64 is sourced from a real-world Internet dermatology specialist hospital diagnosed by dermatologists, and divided into 70%/10%/20% as the train/valid/test sets.

Implementation Details. We employ ResNet50 [20] and BERT [21] as the backbone of the vision and text encoders.

TABLE I
SKIN DISEASE CLASSIFICATION RESULTS ON MSLD AND MSID DATASETS. (UNIT: %)

Init Methods	MSLD				MSID			
	Accuracy	Precision	Recall	F1 score	Accuracy	Precision	Recall	F1 score
ImageNet Init.	81.33±0.69	77.36±1.11	82.00±1.64	79.61±0.67	74.94±0.28	72.59±0.72	67.64±1.73	69.25±1.09
SimCLR [18]	87.56±0.72	86.62±2.72	85.33±1.70	85.92±0.48	83.74±0.32	85.02±0.14	77.72±0.24	80.66±0.13
MoCo [19]	88.15±0.83	89.12±2.54	83.67±1.70	86.26±0.82	84.94±0.13	86.34±0.13	80.45±0.14	82.92±0.03
CLIP [7]	89.56±0.23	86.62±0.64	90.50±1.50	88.51±0.39	81.14±0.25	81.66±0.67	76.95±1.33	78.68±0.73
ConVIRT [13]	82.22±0.12	76.56±0.24	86.50±0.50	81.22±0.09	78.16±0.57	75.91±0.37	73.31±0.79	74.48±0.63
GLoRIA [12]	90.67±0.49	89.90±2.13	89.00±1.45	89.45±0.67	84.18±0.22	86.35±0.16	81.73±0.17	83.68±1.01
M3AE [11]	82.81±1.46	81.30±1.57	79.67±2.62	80.46±1.82	82.53±0.51	81.29±0.37	76.76±1.26	78.58±0.88
M2CG (Ours)	**93.33±0.42**	**92.08±0.69**	**93.00±1.25**	**92.54±0.52**	**85.82±0.32**	**87.61±0.69**	**82.50±0.80**	**84.68±0.75**

TABLE II
CLASSIFICATION RESULTS ON MORE-FINE-GRAINED DERMNET23 AND SKIN64 DATASETS. (ACC@K: TOP-K ACCURACY, UNIT: %)

Init Methods	DermNet23					Skin64				
	Acc@1	Acc@3	Precision	Recall	F1 score	Acc@1	Acc@3	Precision	Recall	F1 score
ImageNet Init.	57.47±0.20	77.23±0.47	52.91±0.15	53.16±0.06	52.64±0.06	58.25±0.14	78.58±0.31	51.06±0.22	47.19±0.23	47.02±0.19
SimCLR [18]	57.08±0.64	77.79±0.57	54.42±0.89	54.88±0.81	54.64±0.95	61.70±0.35	82.38±0.26	50.49±0.23	46.62±0.45	46.83±0.48
MoCo [19]	58.87±0.56	78.54±0.55	54.29±0.51	55.11±1.01	54.29±0.67	62.70±0.29	82.73±0.65	52.35±0.62	48.68±0.28	49.18±0.48
CLIP [7]	59.10±0.03	80.16±0.05	55.50±0.06	53.81±0.15	54.30±0.05	73.59±0.28	90.25±0.29	62.65±1.73	57.13±0.26	57.60±0.41
ConVIRT [13]	58.15±0.89	78.11±0.58	53.89±1.68	53.45±1.04	53.45±1.11	67.01±0.14	85.36±0.45	52.85±0.80	46.84±0.07	46.58±0.12
GLoRIA [12]	59.62±0.43	79.36±0.22	55.43±0.65	56.18±0.65	55.40±0.63	70.38±0.49	89.03±0.05	59.63±0.35	55.69±0.37	56.70±0.15
M3AE [11]	61.87±1.34	82.01±0.86	58.59±2.19	57.84±1.41	57.54±1.85	65.37±0.21	84.21±0.82	59.97±0.35	53.89±1.32	54.87±0.99
M2CG (Ours)	**64.52±0.75**	**82.41±0.23**	**60.82±0.42**	**60.64±0.85**	**60.44±0.72**	**75.07±0.12**	**92.13±0.17**	**63.85±0.01**	**61.71±0.17**	**61.71±0.02**

The multi-modal fusion encoder and the chief complaint generator are constructed based on the Transformer structure [17] with depths of 12-layer, which is a common configuration in prior researches [22]. During finetuning, the pre-trained vision encoder is utilized followed by an additional MLP classifier with 1 hidden layer and ReLU to predict the disease categories. We follow the common practice of updating model parameters end-to-end on the downstream tasks with the labeled training set. The objective is to reduce the cross-entropy loss between the predicted label and the ground truth. All comparison models are trained for 32 epochs during pre-training and 100 epochs during finetuning using AdamW optimizers with a batch size of 32. The learning rates are set to 1e-4 during pre-training and 1e-5 during finetuning, with cosine annealing strategies for learning rate adjustment. All experiments are implemented with PyTorch on Tesla P40. We run each experiment 3 times with distinct random seeds and report the results on average.

B. Comparison Experiments and Discussions

It is important to clarify that as a backbone pre-training approach, M2CG is not opposite but rather collaborative with the works specialized in skin disease diagnosis [4], [5], [23]. Therefore, we evaluate the effectiveness of M2CG in comparison with the pre-training methods, including the ImageNet baseline [24], the popular SSL methods of SimCLR [18] and MoCo [19], and several SOTA VLP methods of CLIP [7], ConVIRT [13], GloRIA [12], and M3AE [11] developed in both natural and medical fields. For fairness, we use a prompt template "Chief complaint: {...}. Consultation dialogue: {...}." as the text input for comparative VLP methods.

The evaluation results are shown in Table I and Table II. An impressive observation is that in most cases the VLP methods (i.e., CLIP, ConVIRT, GLoRIA, M3AE, and our M2CG) competitively outperform the supervised and self-supervised methods, convincingly presenting the benefits of multi-modal semantic knowledge for visual representation. M2CG consistently outperforms the commonly used baseline and SOTA VLP methods by up to 12.00% and 11.11% on the public benchmarks, and 16.82% and 9.7% on the in-house Skin64 dataset in terms of diagnosis accuracy. Specifically, as shown in Table I, our M2CG remarkably surpasses the SOTAs in all evaluation metrics on MSLD and MSID, achieving significant improvements by 12.00% and 10.88% over the commonly used ImageNet baseline, and more than 2.66% and 1.64% over the SOTA VLP methods in diagnosis accuracy. DermNet23 and Skin64 consist of more fine-grained categories of diseases, which are more demanding to capture the discriminative features. In these cases, M2CG consistently outperforms the SOTAs, achieving enhancements of over 2.65% and 1.48% in terms of top-1 accuracy, as shown in Table II. These improvements suggest a remarkable potential of M2CG for fine-grained skin disease diagnosis.

Compared with the SSL methods, M2CG exploits the semantic knowledge from multi-modal data to facilitates better comprehension of lesion images, increasing 13.37% over the SimCLR remarkably in diagnosis accuracy on the extremely fine-grained task of Skin64. Compared to the CLIP and ConVIRT which rely solely on multi-modal contrastive learning, M2CG integrates self-supervised and generative learning within multi-modal contrastive learning and achieves the highest accuracy with enhancements of up to 5.42% and 11.11%. Besides, M2CG employs CMG to fully exploit multi-modal knowledge and concentrate on crucial features, making it more capable of dealing with fine-grained challenges. The trainable parameters of the competitive VLP methods of CLIP, GLoRIA, M3AE, and M2CG are 132M, 128M, 346M, and 260M, respectively. Overall, M2CG is consistently superior

TABLE III
ABLATION STUDY ON SKIN64. (ACC@K: TOP-K ACCURACY, UNIT: %)

SMC	MMC	MMF	CMG	Acc@1	Acc@3	Precision	Recall	F1 score
	✓	✓	✓	70.38	88.47	62.67	56.54	57.55
✓		✓	✓	60.12	81.30	53.54	50.50	50.52
✓	✓		✓	70.71	88.71	60.17	58.33	57.79
✓	✓	✓		68.28	87.03	59.12	55.81	55.01
✓	✓	✓	✓	**75.07**	**92.13**	**63.85**	**61.71**	**61.71**

to the SOTAs with comparable parameters. The outstanding performance demonstrates the effectiveness of M2CG in fine-grained feature extraction and skin disease diagnosis.

C. Ablation Study

To further investigate the key modules (i.e., SMC, MMC, MMF, CMG) of M2CG, we perform a throughout ablation study in Table III. Overall, the results demonstrate all modules contribute to M2CG and the complete M2CG achieves the best results collaboratively. MMC brings the most significant improvements of 14.95% and 10.83% in top-1 and top-3 accuracy, indicating the effectiveness and importance of multi-modal semantic knowledge and multi-modal alignment for lesion image understanding and visual feature learning. Besides, CMG also enhances the ability of M2CG to recognize fine-grained lesions, leading to improvements of 6.79% and 5.1% in top-1 and top-3 accuracy. These improvements suggest the significance and benefits of extracting discriminative features via semantic knowledge to capture fine-grained features with VLP, which is always ignored in coarse-grained works.

IV. CONCLUSION

In this paper, we propose M2CG, a multi-level multi-modal contrastive-generative pre-training approach to enhance the fine-grained feature extraction for skin disease diagnosis. To address the fine-grained challenge of skin disease diagnosis from lesion images, M2CG simultaneously performs feature-level multi-modal contrastive for semantic-guided fine-grained feature learning and semantic-level cross-modal generation to encourage the identification of the key and discriminative features in the lesion images. We collect a large-scale multi-modal skin disease dataset for pre-training. Extensive experiments are conducted on three public benchmarks and a 64-class fine-grained benchmark collected from the real world. The superior performance of M2CG demonstrates that M2CG addresses the fine-grained VLP problem better and is beneficial for multiple downstream skin disease diagnosis tasks.

ACKNOWLEDGMENT

This work was supported in part by the National Key R&D Program of China (2018YFA0704000, 2020AAA0105500), the Beijing Natural Science Foundation (M22024), the National Natural Science Foundation of China (No.62021002, No.61971260), and the Zhejiang Lab Open Research Project (NO. K2022SKI0AB01). This work was also supported by THUIBCS, Tsinghua University, BLBCI, Beijing Municipal Education Commission, and Tsinghua University-Beijing Jingdong Century Trade Co., Ltd. Joint Research Center for Intelligent Retail Technology.

REFERENCES

[1] R. Hay, M. Augustin, C. Griffiths, W. Sterry, and the Board of the International League of Dermatological Societies and the Grand Challenges Consultation groups , “The global challenge for skin health,” *British Journal of Dermatology*, vol. 172, no. 6, pp. 1469–1472, 2015.

[2] F. Liu, Y. Tian, Y. Chen, Y. Liu, V. Belagiannis, and G. Carneiro, “Acpl: Anti-curriculum pseudo-labelling for semi-supervised medical image classification,” in *CVPR*, 2022.

[3] Z. Zhou, L. Qi, X. Yang, D. Ni, and Y. Shi, “Generalizable cross-modality medical image segmentation via style augmentation and dual normalization,” in *CVPR*, 2022.

[4] Y. Lin, Y. Guan, Z. Ma, H. You, X. Cheng, and J. Jiang, “An acne grading framework on face images via skin attention and sfnet,” in *IEEE Conf. Bioinformatics Biomed.*, 2021.

[5] S. Mishra, Y. Zhang, L. Zhang, T. Zhang, X. S. Hu, and D. Z. Chen, “Data-driven deep supervision for skin lesion classification,” in *Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2022.

[6] C. Zhou, M. Sun, L. Chen, A. Cai, and J. Fang, “Few-shot learning framework based on adaptive subspace for skin disease classification,” in *IEEE Conf. Bioinformatics Biomed.*, 2022.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *Proc. Int. Conf. Mach. Learn.*, 2021.

[8] P. Müller, G. Kaissis, C. Zou, and D. Rueckert, “Radiological reports improve pre-training for localized imaging tasks on chest x-rays,” in *Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2022.

[9] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, “Align before fuse: Vision and language representation learning with momentum distillation,” in *NeurIPS*, 2021.

[10] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, “SimVLM: Simple visual language model pretraining with weak supervision,” in *ICLR*, 2022.

[11] Z. Chen, Y. Du, J. Hu, Y. Liu, G. Li, X. Wan, and T.-H. Chang, “Multi-modal masked autoencoders for medical vision-and-language pre-training,” in *Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2022.

[12] S.-C. Huang, L. Shen, M. P. Lungren, and S. Yeung, “Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition,” in *CVPR*, 2021.

[13] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, “Contrastive learning of medical visual representations from paired images and text,” in *Machine Learning for Healthcare Conference*, 2022.

[14] https://www.kaggle.com/datasets/shubhamgoel27/dermnet, kaggle.

[15] S. N. Ali, M. T. Ahmed, J. Paul, T. Jahan, S. M. S. Sani, N. Noor, and T. Hasan, “Monkeypox skin lesion detection using deep learning models: A preliminary feasibility study,” *arXiv preprint arXiv:2207.03342*, 2022.

[16] T. Islam, M. A. Hussain, F. U. H. Chowdhury, and B. M. Riazul Islam, “A web-scraped skin image database of monkeypox, chickenpox, smallpox, cowpox, and measles,” *bioRxiv*, 2022.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *NeurIPS*, 2017.

[18] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proc. Int. Conf. Mach. Learn.*, 2020.

[19] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *CVPR*, 2020.

[20] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *NAACL*, 2019.

[22] C. Li, H. Xu, J. Tian, W. Wang, M. Yan, B. Bi, J. Ye, H. Chen, G. Xu, Z. Cao, J. Zhang, S. Huang, F. Huang, J. Zhou, and L. Si, “mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections,” in *EMNLP*, 2022.

[23] Y. Lin, J. Jiang, Z. Ma, D. Chen, Y. Guan, X. Liu, H. You, J. Yang, and X. Cheng, “Cgpg-gan: An acne lesion inpainting model for boosting downstream diagnosis,” in *IEEE Conf. Bioinformatics Biomed.*, 2022.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *CVPR*, 2009.