Aligning Medical Images with General Knowledge from Large Language Models

Xiao Fang[†1], Yi Lin[†1], Dong Zhang[2], Kwang-Ting Cheng[2], Hao Chen[✉1,3,4]

[1]Department of Computer Science and Engineering, HKUST, Hong Kong, China
[2]Department of Electronic and Computer Engineering, HKUST, Hong Kong, China
[3]Department of Chemical and Biological Engineering, HKUST, Hong Kong, China
[4]HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen, China.
jhc@cse.ust.hk

Abstract. Pre-trained large vision-language models (VLMs) like CLIP have revolutionized visual representation learning using natural language as supervisions, and demonstrated promising generalization ability. In this work, we propose ViP, a novel visual symptom-guided prompt learning framework for medical image analysis, which facilitates general knowledge transfer from CLIP. ViP consists of two key components: a visual symptom generator (VSG) and a dual-prompt network. Specifically, VSG aims to extract explicable visual symptoms from pre-trained large language models, while the dual-prompt network utilizes these visual symptoms to guide the training on two learnable prompt modules, *i.e.*, *context prompt* and *merge prompt*, which effectively adapts our framework to medical image analysis via large VLMs. Extensive experimental results demonstrate that ViP can outperform state-of-the-art methods on two challenging datasets. The code is available at https://github.com/xiaofang007/ViP.

Keywords: Prompt Learning · Vision-Language Models · Large Language Model · Medical Image Analysis.

1 Introduction

Medical image analysis plays a crucial role in healthcare, enabling non-invasive diagnosis and treatment of various medical conditions [14,15,3,23]. With the advent of deep learning techniques, computer-aided medical image analysis has achieved remarkable success in numerous scenarios. Current methods generally adopt the supervised learning paradigm which requires a large amount of labeled data for model training. However, this paradigm relies on manual annotation of medical images, which is time-consuming and labor-intensive [24].

The emergence of large vision language models (VLMs) [11,12,13] makes it possible to transfer knowledge from large-scale pre-trained models to task-specific medical image analysis models with limited data. One prominent example is Contrastive Language-Image Pre-training (CLIP) [21], which is pre-trained

---

† Equal contribution; ✉ corresponding author.

on 400 million image-text pairs using contrastive learning. In detail, it comprises a vision and a text encoder that encodes an image and its corresponding text snippet into visual and textual embeddings, respectively. While CLIP has demonstrated great potential in transfer learning across diverse tasks in universal scenes, its direct applications to the medical domain raise challenges. This is because CLIP is pre-trained mainly on web-scraped data, which primarily comprises natural image-text pairs and lacks medical data due to privacy concerns, while the category texts of medical images tend to be abstract medical lexicons, which can be hard for CLIP to interpret. Inspired by recent work [18,20], we propose to address the interpreting challenge by translating abstract medical lexicons to visual symptoms that are shared across natural and medical domains, such as color, shape and texture. In this way, VLMs can learn to align image features with visual features that are easily interpreted. This process also aligns with the diagnostic approach employed by medical professionals, who diagnose diseases based on related visual features observed in medical images.

In this paper, we propose ViP, a novel **Vi**sual symptom-guided **P**rompt learning framework that promotes general knowledge transfer of CLIP [21]. The framework consists of two main components: A visual symptom generator (VSG) and a dual-prompt network. VSG queries pre-trained large language models (LLMs) to generate visual symptoms, which serve as text inputs for the dual-prompt network. The dual-prompt network enhances the generalization ability of CLIP by training two learnable prompt modules: *context prompt* (CoP) and *merge prompt* (MeP). CoP refines visual symptoms by incorporating medical task context while MeP aggregates text features of visual symptoms. The proposed framework is evaluated on two public datasets, including Pneumonia [9] and Derm7pt [8]. Extensive experimental results demonstrate that ViP outperforms state-of-the-art methods, highlighting the efficacy of each component in our framework.

The main contributions of our work are as follows: 1) We reveal the significant impact of LLMs on prompt engineering, showcasing their influence on enhancing interpretability and performance. 2) We propose ViP that leverages LLMs to generate visual symptoms in a scalable manner and employs two learnable prompt modules to facilitate knowledge transfer from CLIP to the medical domain. 3) We conduct extensive experiments on two datasets, and the result demonstrates the strong generalization ability of ViP to medical image analysis.

2 Method

2.1 Overall Pipeline

The pipeline of our method is presented in Fig. 1. We consider an input image x and a set of disease labels $C = \{c_1, c_2, ..., c_n\}$, where we denote N as the total number of disease categories, with $N = n$. The process begins by passing x through a pre-trained vision encoder in the dual-prompt network to compute a
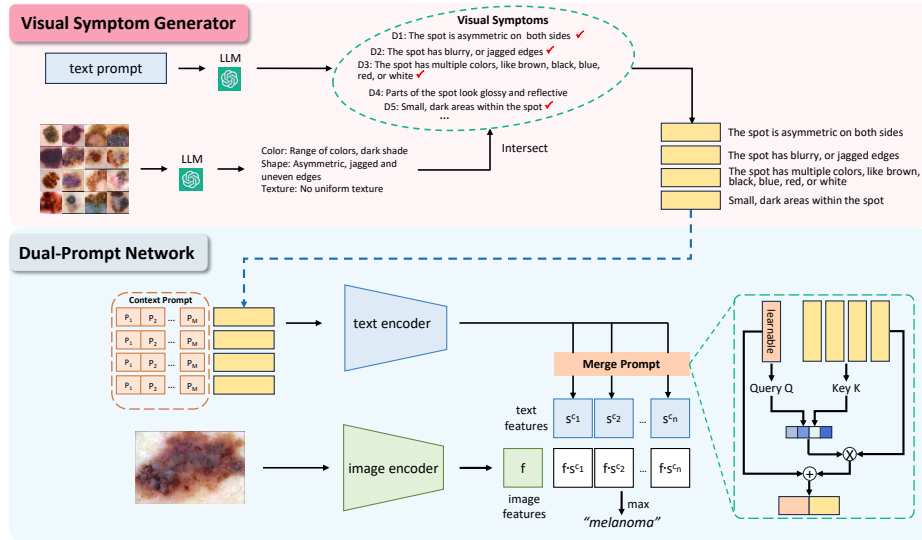
Fig. 1: Overview of ViP, which consists of a visual symptom generator (VSG) and a dual-prompt network. The visual symptoms predicted by VSG are used as inputs for downstream networks (marked by the blue dashed line).

feature vector f. In parallel, several visual symptoms are generated by the visual symptom generator (VSG) for each disease category. These visual symptoms then undergo transformation in the context prompt module (CoP) to create textual input embeddings for the dual-prompt network. These textual embeddings are then processed through the pre-trained text encoder to compute the textual features for each visual symptom. Next, the merge prompt module (MeP) aggregates text features to obtain a representative feature s^c for disease category c. Going over all categories $c \in C$, we obtain a set of aggregated visual descriptive features $S = \{s^{c_1}, s^{c_2}, ..., s^{c_n}\}$. Finally, we predict the disease category with the highest cosine similarity score $f \cdot s^c, c \in C$. In the following sections, we will explain the VSG and the dual-prompt network in detail.

2.2   Visual Symptom Generator (VSG)

VSG aims to generate a comprehensive set of visual symptoms specific to each disease category. Impressed by the broad knowledge possessed by LLMs and that they can be easily queried with natural language, we propose a two-stage process to construct this set by prompting a large language model, such as GPT-4 [1]. First, we use a text-only prompt to obtain a coarse set of visual symptoms. We prompt the language model with the following text as the input:

```
Q: I am going to use CLIP, a vision-language model to detect {category}
   in {modality}. What are useful medical visual features for diagnosing
```

```
{category}? Please list in bullet points and explain in plain words
that CLIP understands. Avoid using words such as {category}.
```

where {category} is substituted for a given category $c \in C$ and {modality} is substituted for the imaging modality of the dataset, *e.g.*, dermoscopic images. The prompt is designed to provide sufficient background for GPT-4 and ensure the answers are understandable by CLIP. Next, we refine the coarse set by leveraging the visual-question-answering function of GPT-4. We prompt it with multiple images for each disease category using the following query:

```
Q: Please provide visual features regarding color, shape, and texture of
this {category} image, which contains 16 sub-images.
```

After receiving the response that encompasses a set of commonly observed visual features across images, the refined set is obtained by intersecting the initial coarse set with the response. Fig. 2 demonstrates the visual symptoms generated by GPT-4 [1] using our designed pipeline. As expected, generated visual symptoms typically cover descriptions of color and shape of lesions, presence or absence of certain structures, and other relevant visual features.

2.3   Dual-Prompt Network

The dual-prompt network is built upon CLIP. We freeze the image encoder and text encoder of CLIP to retain the general knowledge from the large-scale pre-training data. Unlike conventional CLIP-based approaches that rely on category names for textual input, we use visual symptoms generated from the VSG to enable the model to facilitate the alignment of image features with visual descriptive features. However, the generalization ability of our framework is still limited. This limitation arises due to potential deviations from the expected CLIP text input format in the response from LLMs, and the inherent challenge of effectively aggregating visual symptoms into a disease representation without explicit training [17,6]. Therefore, we further propose two learnable prompt modules: *context prompt* (CoP) and *merge prompt* (MeP), to enhance the model generalization ability.

CoP. In addition to category names, context words help to form a complete sentence that specifies the context of the image, which plays a crucial role in the textual input of CLIP. For example, CLIP prepends category names with the context {a photo of a}. Similarly, it is desirable to prepend visual symptoms with a customized template to capture the context of medical tasks. However, it is challenging to design hand-crafted templates for visual symptoms due to their more complex phrase structure. Motivated by [26], we introduce a set of learnable tokens $\{p_i\}_{i=1}^{M}$, where $p_i \in \mathbb{R}^d, i = 1, 2, ...M$, and d is the text embedding dimension, before visual symptoms to automatically learn the context of medical tasks in a data-driven manner. Specifically, given a category $c \in C$, and a visual symptom word embedding e_d, the final textual input word embedding T for the text encoder is the concatenation of the learnable tokens and e_d, which can be formulated as $T = \text{Concat}(p_1, p_2...p_M, e_d)$.
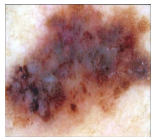
Pneumonia	Normal Lung	Melanoma	Nevus
Presence of pleural effusion	Normal Shape and size	The spot is asymmetric on both sides	The spot is symmetric on both sides
Presence of cavitation	Sharp and clear lung borders	The spot has blurry, or jagged edges	The spot has clear, well-defined edges
Presence of consolidation	Presence of lung markings	The spot has multiple colors, like brown, black, blue, red, or white	The spot has a consistent color, usually brown, tan or black
Air bronchogram sign	Absence of focal opacities	Small, dark areas within the spot	Redness or swelling around the spot

Fig. 2: Example visual symptoms generated by GPT-4 [1].

MeP. After processing visual symptoms via text encoder, the next step is to merge visual symptoms into a single representation. Previous methods [16,18,2] adopt the average function, which treats all visual symptoms as equally important, or the max function, which diagnoses based on the most prominent feature. However, these functions suffer from inherent bias because not all visual symptoms contribute equally to a disease. Additionally, it is impossible to accurately diagnose a disease based solely on the most prominent visual symptom in all cases. Therefore, we introduce a learnable token for each disease category to learn the representative feature of the disease. Specifically, given a category $c \in C$, text features matrix $T = [T_1^c, T_2^c, ..., T_k^c]^T$, where $T \in \mathbb{R}^{k \times d}$ and d is the text embedding dimension, which is obtained by processing related visual symptoms through the text encoder, and a learnable grouping token $g \in \mathbb{R}^d$, we first project g and T into query $Q \in \mathbb{R}^d$ and key $K \in \mathbb{R}^{k \times d}$ with different weights $W_q \in \mathbb{R}^{d \times d}$ and $W_k \in \mathbb{R}^{d \times d}$, which can be formulated as:

$$Q = gW_q, K = TW_k. \tag{1}$$

The aggregated feature s^c is calculated by combining the grouping prompt g and weighted text features matrix T, which can be formulated as:

$$s^c = g + \text{Softmax}(\frac{QK^T}{\sqrt{d}})T. \tag{2}$$

After obtaining the aggregated visual descriptive features of all disease categories, CoP and MeP are jointly optimized with a cross-entropy loss, which can be formulated as:

$$L_{ce} = -\log \frac{\exp(f \cdot s^{c_y}/\gamma)}{\sum_{i=1}^{N} \exp(f \cdot s^{c_i}/\gamma)}, \tag{3}$$

where c_y denotes the ground truth disease category and γ is a learned temperature.
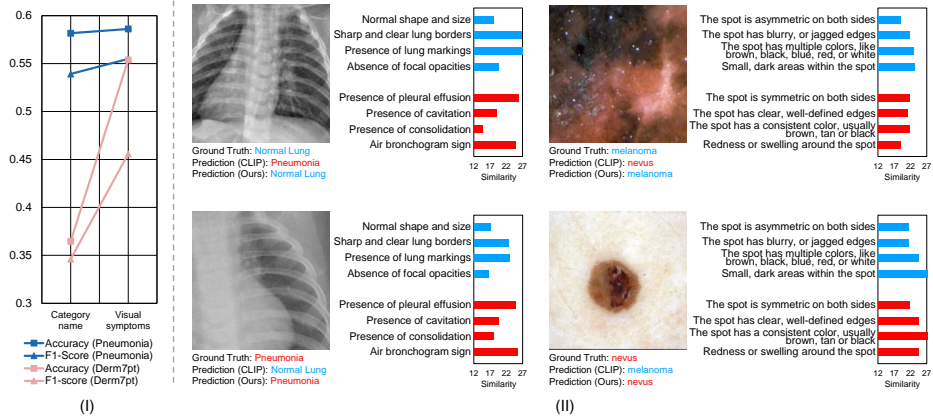
Fig. 3: (I) Zero-shot CLIP with category name or visual symptoms as text inputs. (II) Diagnostic process based on cosine similarity scores between images and visual symptoms.
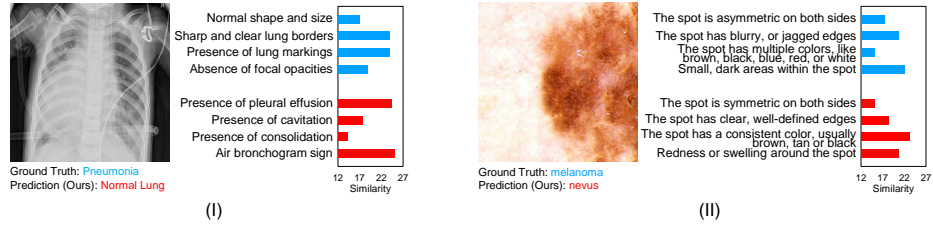

Fig. 4: Failure cases in the zero-shot experiment.

3   Experiments

3.1   Dataset and Implementation Details

Dataset. We conduct experiments on two publicly available datasets: Pneumonia [9] and Derm7pt [8]. Pneumonia consists of chest X-ray images categorized as either normal lung or pneumonia. The official split of this dataset contains 5232 images for training and 624 images for testing. We further randomly divide the training set with a ratio of 9:1 for training and validation. Derm7pt consists of over 2000 clinical and dermoscopic images. Following [19], we filter the dataset to obtain 827 images belonging to the "melanoma" and "nevus" classes, and split the dataset into 346, 161, and 320 images for training, validation, and testing, respectively. For both datasets, we adopt Accuracy (ACC) and Macro F1-score (F1) as evaluation metrics. Macro F1-score addresses the data imbalance issue by computing the arithmetic mean of all per-class F1 scores.

Implementation Details. We train the proposed ViP model on an NVIDIA RTX 3090 GPU. Throughout the experiments, we average the results of three

Table 1: Result comparisons with SOTAs. The mean and standard deviation is computed across three vision backbones.

Method	Pneumonia		Derm7pt	
	ACC	F1	ACC	F1
CoOp [26]	$0.8337_{0.019}$	$0.8148_{0.017}$	$0.7823_{0.005}$	$0.7328_{0.017}$
CoCoOp [25]	$0.8440_{0.025}$	$0.8217_{0.032}$	$0.7668_{0.014}$	$0.6647_{0.057}$
KgCoOp [22]	$0.8303_{0.022}$	$0.8010_{0.027}$	$0.7726_{0.009}$	$0.7093_{0.033}$
Bayesian [4]	$0.8301_{0.041}$	$0.8081_{0.048}$	$0.6921_{0.014}$	$0.5561_{0.054}$
MaPLe [10]	$0.8553_{0.034}$	$0.8393_{0.036}$	$0.7903_{0.038}$	$0.7250_{0.073}$
Supervised	$0.8660_{0.025}$	$0.8530_{0.025}$	$0.7277_{0.044}$	$0.6236_{0.093}$
ViP_{ours}	$\mathbf{0.8669}_{0.031}$	$\mathbf{0.8494}_{0.036}$	$\mathbf{0.8111}_{0.007}$	$\mathbf{0.7730}_{0.015}$

vision backbones in CLIP [21], *i.e.*, ViT-B/16 [5], ViT-L/14 [5], and ResNet-50 [7]. We follow CLIP [21] to set the text embedding dimension d to 512. We follow CoOp [26] to learn a unified task context and set the length M of the context prompt (CoP) to 4. Training is done with SGD and an initial learning rate of 0.001. The training epoch is set to 50. We follow CLIP [21] to set the temperature γ in the cross-entropy loss to $\frac{1}{100}$.

3.2   Comparisons with State-of-the-art Methods

Effectiveness of explainable visual symptoms. We conduct a zero-shot experiment to evaluate the effectiveness of visual symptoms for disease diagnosis, while also providing explanations for the decisions. Specifically, our approach makes decision by comparing images to the average embedding of visual descriptive features. As shown in Fig. 3(I), compared with zero-shot CLIP, our method achieves 0.44% and 18.73% accuracy improvement, and F1-score gains of 1.58% and 10.98% over Pneumonia [9] and Derm7pt [8], respectively. This suggests that LLMs can provide useful knowledge for the medical domain. We further analyzed cases where our method correctly predicts the disease category while CLIP fails, as shown in Fig. 3(II). Our framework improves diagnosis accuracy due to the relatively higher similarity between the images and the characteristics of the correct category. For instance, Fig. 3(II)(d) is diagnosed as nevus because it demonstrates higher similarity with several characteristics of nevus such as clear edges, consistent brown color, and swelling around the lesion, despite there being a small, black area inside the lesion. However, as shown in Fig. 4, there are instances where our method fails to predict the disease category. In Fig. 4(I), although the image exhibits higher similarity with pneumonia characteristics, such as the presence of pleural effusion and air bronchogram sign, the average similarity is lower due to the less obvious symptoms of cavitation and consolidation. This highlights the limitation of using the average function to represent the overall visual features of a disease. In Fig. 4(II), our method fails to diagnose

Table 2: Ablation study results. "Context" and "Merge" denote context prompt (CoP) and merge prompt (MeP), respectively. "Max" and "Mean" denote the maximum and average of visual descriptive features, respectively.

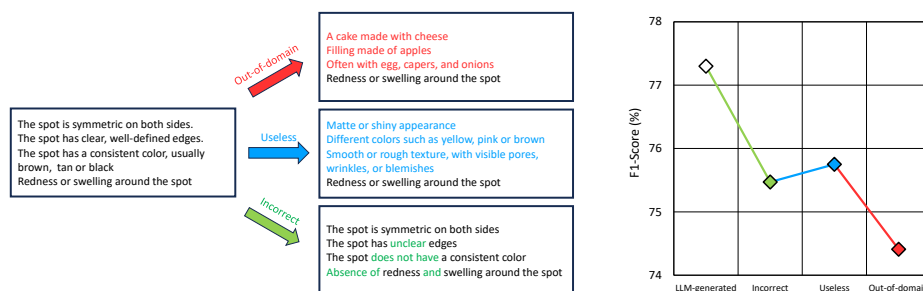| Context	Merge	Pneumonia		Derm7pt	
		ACC	F1	ACC	F1
✗	✗	0.5861	0.5549	0.5539	0.4558
✗	✓	0.8486	0.8312	0.7531	0.6194
✓	Max	0.8390	0.8223	0.7970	0.7506
✓	Mean	0.8550	0.8347	0.8041	0.7646
✓	✓	**0.8669**	**0.8494**	**0.8111**	**0.7730**

Fig. 5: Ablation study comparing with different types of knowledge.

correctly because the image shares high similarity with nevus characteristics, such as brown color.

Comparison with related methods. We further compare ViP with several SOTA prompt-based models to evaluate the generalization ability. As shown in Table 1, ViP achieves highest accuracy of 86.69%, 81.11%, and F1-score of 84.94% , 77.3% on Pneumonia [9] and Derm7pt [8], respectively, indicating the strong generalization ability of our method. Moreover, compared with the fully supervised learning mode, ViP achieves competitive result on Pneumonia [9] , but outperforms a great margin on Derm7pt [8] where there is less training data, demonstrating the strong generalization ability of ViP in low-resource settings.

3.3   Ablation Study

Effectiveness of each component. We conduct ablation studies to explore the effectiveness of each component in ViP, as shown in Table 2. Compared with zero-shot baseline, both the integration of CoP and MeP exhibit considerable improvement, demonstrating the importance of learning medical task context and effective aggregation of visual symptoms. Moreover, compared with non-parametric aggregation methods, such as average and max functions [2,18], our proposed MeP outperforms in both datasets. This result further validates the effectiveness of our method.

Knowledge Faithfulness. We conduct an additional experiment to validate our argument that LLM-generated visual symptoms provide useful knowledge for the generalization to the medical domain. As shown in Fig. 5, we replace the visual symptoms of nevus with three types of knowledge: 1) Out-of-domain knowledge, involving visual symptoms unrelated to the medical domain, such as descriptions of food. 2) Useless knowledge, referring to descriptions associated with our target disease but do not provide useful information for diagnosis, such as descriptions of skin structure. 3) Incorrect knowledge, which provides erroneous symptoms for diagnosis. In this experiment, we alter certain words in the descriptions to their antonyms to create misleading descriptions of nevus. Compared to other variations, LLM-generated knowledge achieves best performance, indicating that accurate visual symptoms contribute to the generalization in the medical domain.

4 Conclusion

This paper presented a novel visual symptom-guided prompt learning pipeline, referred to as ViP, which effectively transfers knowledge from VLMs to medical image analysis. By leveraging pre-trained LLMs, ViP generates useful visual symptoms to guide CLIP in aligning image features with visual symptoms. Additionally, ViP incorporates two learnable prompt modules, context prompt and merge prompt, to further enhance the generalization ability. Experimental results underscored the effectiveness of each module and the superior performance of our pipeline to state-of-the-art methods. Future work will focus on extending the framework to other medical image analysis tasks, such as the diagnosis of rare diseases and malformed organs, where data and annotations are scarce and costly. Additionally, we will investigate techniques to enhance the interpretability of context prompt.

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

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv (2023)
2. Byra, M., Rachmadi, M.F., Skibbe, H.: Few-shot medical image classification with simple shape and texture text descriptors using vision-language models. arXiv (2023)

3. Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y.: Transmorph: Transformer for unsupervised medical image registration. Medical image analysis **82**, 102615 (2022)

4. Derakhshani, M.M., Sanchez, E., Bulat, A., da Costa, V.G.T., Snoek, C.G., Tzimiropoulos, G., Martinez, B.: Bayesian prompt learning for image-language model generalization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15237–15246 (2023)

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. In: International Conference on Learning Representations (2020)

6. Franquet, T.: Imaging of pneumonia: trends and algorithms. European Respiratory Journal **18**(1), 196–208 (2001)

7. 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)

8. Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G.: Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE journal of biomedical and health informatics **23**(2), 538–546 (2018)

9. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. cell **172**(5), 1122–1131 (2018)

10. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19113–19122 (2023)

11. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)

12. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10965–10975 (2022)

13. Li, Y., Fan, H., Hu, R., Feichtenhofer, C., He, K.: Scaling language-image pre-training via masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23390–23400 (2023)

14. Lin, Y., Fang, X., Zhang, D., Cheng, K., Chen, H.: Boosting convolution with efficient mlp-permutation for volumetric medical image segmentation. arXiv (2023)

15. Lin, Y., Zhang, D., Fang, X., Chen, Y., Cheng, K.T., Chen, H.: Rethinking boundary detection in deep learning models for medical image segmentation. In: International Conference on Information Processing in Medical Imaging. pp. 730–742. Springer (2023)

16. Liu, J., Hu, T., Zhang, Y., Gai, X., Feng, Y., Liu, Z.: A chatgpt aided explainable framework for zero-shot medical image diagnosis. arXiv (2023)

17. Markovic, S.N., Erickson, L.A., Rao, R.D., McWilliams, R.R., Kottschade, L.A., Creagan, E.T., Weenig, R.H., Hand, J.L., Pittelkow, M.R., Pockaj, B.A., et al.: Malignant melanoma in the 21st century, part 1: epidemiology, risk factors, screening, prevention, and diagnosis. In: Mayo Clinic Proceedings. vol. 82, pp. 364–380. Elsevier (2007)

18. Menon, S., Vondrick, C.: Visual classification via description from large language models. In: The Eleventh International Conference on Learning Representations (2022)

19. Patrício, C., Neves, J.C., Teixeira, L.F.: Coherent concept-based explanations in medical image and its application to skin lesion diagnosis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3798–3807 (2023)
20. Qin, Z., Yi, H.H., Lao, Q., Li, K.: Medical image understanding with pretrained vision language models: A comprehensive study. In: The Eleventh International Conference on Learning Representations (2022)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
22. Yao, H., Zhang, R., Xu, C.: Visual-language prompt tuning with knowledge-guided context optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6757–6767 (2023)
23. You, D., Liu, F., Ge, S., Xie, X., Zhang, J., Wu, X.: Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. pp. 72–82. Springer (2021)
24. Zheng, F., Cao, J., Yu, W., Chen, Z., Xiao, N., Lu, Y.: Exploring low-resource medical image classification with weakly supervised prompt learning. Pattern Recognition **149**, 110250 (2024)
25. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825 (2022)
26. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)