Disease-informed Adaptation of Vision-Language Models

Jiajin Zhang[1], Ge Wang[1], Mannudeep K. Kalra[2], and Pingkun Yan[1]

[1] Department of Biomedical Engineering and Center for Biotechnology and Interdisciplinary Studies, Rensselaer Polytechnic Institute, Troy, NY, USA
[2] Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

Abstract. In medical image analysis, the expertise scarcity and the high cost of data annotation limits the development of large artificial intelligence models. This paper investigates the potential of transfer learning with pre-trained vision-language models (VLMs) in this domain. Currently, VLMs still struggle to transfer to the *underrepresented* diseases with minimal presence and *new* diseases entirely absent from the pretraining dataset. We argue that effective adaptation of VLMs hinges on the nuanced representation learning of disease concepts. By capitalizing on the joint visual-linguistic capabilities of VLMs, we introduce disease-informed contextual prompting in a novel disease prototype learning framework. This approach enables VLMs to grasp the concepts of new disease effectively and efficiently, even with limited data. Extensive experiments across multiple image modalities showcase notable enhancements in performance compared to existing techniques. The code will be made publicly available on our github page: https://github.com/RPIDIAL/Disease-informed-VLM-Adaptation.

Keywords: VLM · Transfer Learning · Underrepresented/New Diseases.

1 Introduction

In medical image analysis, the scarcity of required expertise and the high cost of data annotation hinder the development of deep learning models [15, 16, 20, 24–28]. Thus, transfer learning of pre-trained models emerged as a practical solution. Recent advancements have seen pre-trained vision language models (VLMs), such as the CLIP-based networks [2, 7, 11, 19], achieved impressive domain adaptation capabilities. The integration of natural language processing allows VLMs to learn better visual representations with an appropriate alignment with textual concepts, thereby facilitating model generalizability. However, the recent works [18, 21] showed that the efficacy of VLMs diminishes when the relevant category are only sparsely represented in the training data. This paper attempts to address two major challenges in adapting VLMs in computer-assisted diagnosis with medical images, including *underrepresented diseases* with minimal presence and *new diseases* entirely absent from the pretraining dataset.

In a clinical environment, it is often feasible to obtain a limited number of data related to an emerging or different disease for model adaptation. For example, during the early stage of the COVID-19 pandemic, a selection of image samples were rapidly shared to spread awareness and understanding of the disease [22]. Consequently, employing data-efficient methods for adapting pre-trained large models emerges as a viable path for medical image analysis. Several methods have been proposed for transfer learning with VLMs. Adapter-based approaches like linear probing [19], CLIP-Adapter [5], and Tip-Adapter [29] adopt an classifier to tailor pre-trained visual encoders on new tasks, by only tweaking the final layers. Alternatively, prompting-based methods like CoOp [31] and CoCoOp [30] focus on optimizing learnable prompts without actual text inputs, prioritizing performance over the acquisition of meaningful concepts. The recent method MaPLe [13] utilizes dual-modality prompts to adapt CLIP to new domains, yet it demands extensive data for finetuning and may not suitable for diseases with limited data. In this paper, we propose a new framework of disease-informed adaptation to address the above limitations.

Our research contends that effective representation learning of disease concepts is central to successfully adapting vision language models (VLMs). Our method innovates in the following two fronts. **1)** *Disease-informed Contextual Prompting* (DiCoP) utilizes the clinical knowledge to craft prompts for representing the concepts of the target diseases. The prompts highlight the disease characteristics with descriptive attributes, such as texture, shape, and location. To overcome the problem that crafted prompts lack instance specifics, we further propose to enhance the textual prompts with image context features. **2)** *Disease Prototype Learning* (DPL) addresses the lack of structural regulation in the latent space of CLIP-based VLMs [6], which is critical for recognizing the target diseases. The DPL framework fine-tunes the image encoder to actively learn the prototypes of diseases, and regularizes the geometric structure of the learned representations for the downstream visual recognition tasks.

To demonstrate the effectiveness and efficiency of our proposed approach, we carried out extensive experiments across various imaging modalities and diagnostic tasks. Our evaluation specifically targets two illustrative use cases. In the first scenario, focused on identifying cancerous pathological images, the target cancer type is *underrepresented* due to much less data compared to other cancer types in the pretraining dataset. The second scenario mirrors a clinical situation, where a *newly identified* disease was not present in the pretraining dataset with only a few samples available for model adaptation. An example of this type is the COVID-19 pandemic's onset. Our method can help improve the preparedness and response when we deal with future health crisis.

2   Method

Our method utilizes medical VLMs, pretrained in the CLIP style, to diagnose underrepresented or new diseases. The core idea of our method is the effective representation learning of disease concepts. This involves two steps. First,
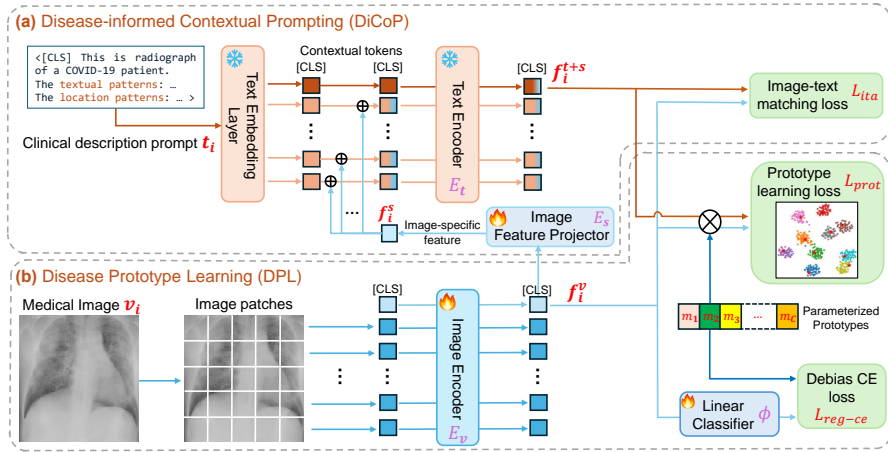
Fig. 1. The framework overview. **(a)** DiCoP produces prompts informed by specific diseases. **(b)** DPL enables learning disease representations with limited data.

Disease-informed Contextual Prompting (DiCoP) is designed to bridge the concepts of new diseases with established clinical knowledge and instance-specific features. Then, *Disease Prototype Learning* (DPL) is utilized to learn disease prototypes by imposing geometric regularization on the learned representations.

As shown in Fig. 1, a CLIP model comprises a text encoder E_t and an image encoder E_v. We denote a labeled downstream dataset of target diseases as $\{(v_i, y_i)\}_{i=1}^N$. The images v_i are grouped into C categories, each represented by a distinct disease or condition. Our proposed DiCoP method utilizes clinical knowledge to generate text prompts t_i for each image. The image and text encoders, E_v and E_t, process each (v_i, t_i) pair, using the [CLS] token to derive global visual and linguistic representations $f_i^t, f_i^v \in \mathbb{R}^{h \times 1}$, respectively. We **freeze** the pretrained text encoder to prevent it from overfitting to the small set of hand-crafted prompts.

2.1 Disease-informed Contextual Prompting (DiCoP)

Drawing on insights from the prior research on natural images analysis [14, 18, 21, 30], we underscore the value of incorporating both *clinical knowledge* and *image-specific features* into contextual prompts to enhance the transferability of VLMs. The key lies in linking concepts of underrepresented/new diseases with the existing clinical knowledge. For instance, radiological descriptions of texture attributes can aid in illustrating pneumonia. Through this approach, the concept of pneumonia is connected to the pre-existing radiology corpus via linguistic semantics by the text encoder E_t. To this end, we propose crafting contextual prompts containing notable attributes of target categories. In particular, we

develop descriptive contextual prompts using the template

$$\texttt{Prompt}_k|_{k=1}^{C} = \texttt{Des}_k(\texttt{texture}) \oplus \texttt{Des}_k(\texttt{location}) \oplus \texttt{Des}_k(\texttt{shape}), \qquad (1)$$

where $\oplus$ symbolizes the concatenation of descriptions $\texttt{Des}_k(\cdot)$ for the k_{th} disease category, covering each attribute (texture, location, shape). Each description $\texttt{Des}_k(\cdot)$ is manually designed with clinical knowledge. The predefined prompts are fixed for all samples of the same category, $i.e.$, $t_i = \texttt{Prompt}_k$ if $i \in S_k$.

Given the visual diversity of images within the same category, we follow CoCoOp [30] to enrich text-only prompts in Eq. 1 with image-specific features. Fusing general clinical knowledge and specific image variability leads to accurate contextual prompts. As depicted in Fig. 1, an image feature projector E_s is introduced to extract image-specific representations $f_i^s = E_s(f_i^v)$, where f_i^v is the visual [CLS] token encoded by E_v. The dimensionality of image-specific representation f_i^s matches the contextual token embeddings. f_i^s is then added to each contextual token embeddings, except for the [CLS] token, to enhance the prompts with image-specific information. We denote the representation of text-only prompts as $\{f_k^t\}_{k=1}^{C}$, and the representations augmented with image-specific features as $\{f_i^{t+s}\}_{i=1}^{N}$. An image-text alignment loss L_{ita} is defined as follows to encourage the model to maximize the cosine similarity between the matched images and prompts, while reducing the similarity between unmatched pairs. τ_1 is a temperature hyperparameter.

$$L_{ita} = \frac{1}{2N} \sum_{i=1}^{N} \left[\frac{\exp(f_i^v \cdot f_i^{t+s}/\tau_1)}{\log\left(\sum_{j=1}^{N} \exp(f_j^v \cdot f_i^{t+s}/\tau_1)\right)} + \frac{\exp(f_i^v \cdot f_i^{t+s}/\tau_1)}{\log\left(\sum_{j=1}^{N} \exp(f_i^v \cdot f_j^{t+s}/\tau_1)\right)} \right] \qquad (2)$$

2.2   Disease Prototype Learning (DPL)

By employing DiCoP generated prompts, which combines clinical knowledge and image-specific feature, alongside a limited number of new image samples, we fine-tune medical VLMs to learn representations of underrepresented or new diseases. However, the original CLIP-based VLMs do not impose any geometric constraints on the representations within each modality, which is a limitation [6] for transferring VLMs to the downstream disease diagnosis tasks. To address the challenge, we propose a framework that enhances the explicit representation learning for the disease of each category.

More specifically, we employ trainable parameters $\{m_k\}_{i=1}^{C}$ to represent the C distinct disease prototypes. Each prototype m_k is initialized by f_k^t, which is the representation of the text-only $\texttt{Prompt}_k$ without image-specific features. Following Eq. 3, we fine-tune the model and the prototypes simultaneously by minimizing the cosine similarity for samples grouped under the same category and their respective prototype m_k. Concurrently, we maximize the separability between different disease prototypes.

$$L_{prot} = \sum_{k=1}^{c} \frac{1}{|2S_k|} \sum_{i \in S_k} \left[\exp\left(\frac{f_i^v \cdot m_k}{\tau_2}\right) + \exp\left(\frac{f_i^{t+s} \cdot m_k}{\tau_2}\right) \right] - \lambda_1 \sum_{k \neq j} \exp\left(\frac{m_k \cdot m_j}{\tau_2}\right) \quad (3)$$

where τ_2 is another temperature hyperparameter and λ_1 is a weighting factor.

As in Fig. 1, the linear classifier ϕ classifies an image with prediction probabilities $\boldsymbol{p}_i = \phi(\boldsymbol{f}_i^v)$. To avoid overfitting the learned disease prototypes $\boldsymbol{m}_k$ to the scarce samples, we further employ the disease-informed prompts, specified in Eq. 1, as concept anchors to regularize the $\boldsymbol{m}_k$. Subsequently, we incorporate a regularized cross-entropy loss as

$$L_{reg-ce} = -\frac{1}{N}\sum_{i=1}^{N}\log(\boldsymbol{p}_i)\cdot\boldsymbol{y}_i + \frac{\lambda_2}{C}\sum_{k=1}^{C}||\boldsymbol{m}_k - \boldsymbol{f}_k^t||_2 \ , \tag{4}$$

where λ_2 is a weighting factor. The regularization term avoids excessive model adaptation to the limited samples by preserving a minimal ℓ_2-norm distance between the disease prototypes $\boldsymbol{m}_k$ and the representations $\boldsymbol{f}_k^t$ of the disease-informed textual prompts. This helps align the model's learning closer to clinical knowledge rather than being narrowly tailored to the limited data. $\boldsymbol{E}_v$, $\boldsymbol{E}_s$, ϕ and $\{\boldsymbol{m}_k\}_{i=1}^C$ are optimized simultaneously by minimizing the loss $L_{total} = L_{ita} + L_{prot} + L_{reg-ce}$. Only the vision branch $\boldsymbol{E}_v$ and ϕ are used for the inference.

3   Experiments and Results

We evaluated the proposed method on two datasets, PanNuke and COVID-x. **PanNuke** [4] contains 7,558 pathological images for 19 organs. The task is a binary classification (malignant or benign) based on the tissue appearances. When applied to downstream tasks, PLIP showed suboptimal zero-shot performance (F1<0.5) on certain organs, including the kidney, colon, and pancreas, because of their *under-representation* in the pretraining dataset [11]. We followed PLIP [11] to resize images to 224×224 pixels, and divided the data into training/validation/test sets with the ratio of $7 : 1 : 2$. **COVID-x**(v6) [23] contains chest x-ray images collected from multiple countries. The task is a 3-way classification (COVID-19, non-COVID pneumonia, or normal). BioViL [2], pretrained on the MIMIC dataset [12], was developed prior to the COVID pandemic. COVID-19 is thus a *new* disease for the model. Since we do not have access to the test set, we split the original training set of 29,634 images into train and validation subsets with a ratio of $7 : 1$, and used the original validation set of 400 images for test. We followed BioViL to resize all images to 360×360 pixels.

3.1   Implementation Details

Network Architectures: PLIP uses ViT-B/16 [3] as its image encoder, while BioViL utilizes ResNet-50 [8].[3] The image feature projector $\boldsymbol{E}_s$ employs a two-layer bottleneck architecture (Linear-ReLU-Linear), compressing the input dimension by a factor of 16. The classifier consists of a simple linear layer, mapping the image features to K classes (class number $K \le$ prototype number C).

---

[3] The text encoders are frozen. Since it is computationally expensive to directly tune the vision encoders, we adopted LoRA [10] (rank $r = 4$ and scaling $\alpha = 4$) to fine-tune PLIP on PanNuke. The BioViL was finetuned on COVID-x using layer-wise learning rate decay [9] (decay factor $\beta = 0.9$).

Table 1. 10-run average AUC and F1 scores of models on PanNuke with 5% training data. The best performance is in **bold** and the second best is <u>underlined</u>.

Method	Kidney		Colon		Pancreatic		All Organs	
	F1	AUC	F1	AUC	F1	AUC	Weighted F1	AUC
Linear Probing	0.560^*	0.660^*	0.748^*	0.794^*	0.810^*	0.754^*	0.740^*	0.779^*
CLIP-Adapter	0.587^*	0.703^*	0.774^*	0.853^*	0.853^*	0.812^*	0.786^*	0.825^*
Tip-Adapter	0.610^*	0.722^*	0.788^*	0.859^*	0.877^*	0.830^*	0.802^*	0.843^*
CoOp	0.617^*	0.730^*	0.810^*	0.863^*	0.874^*	0.828^*	0.804^*	0.847^*
CoCoOp	<u>0.647^*</u>	<u>0.753^*</u>	0.826^*	0.882^*	<u>0.903^*</u>	<u>0.861</u>	<u>0.838^*</u>	<u>0.880^*</u>
MaPLe	0.635^*	0.745^*	<u>0.836^*</u>	<u>0.891</u>	0.891^*	0.853^*	0.827^*	0.869^*
DiCoP + DPL (ours)	**0.692**	**0.786**	**0.860**	**0.897**	**0.920**	**0.875**	**0.865**	**0.896**

* $p < 0.05$ in the one-tailed paired samples Student's t-test with 10-run results using our method.

Hyperparameters and training: All models and prototypes are trained for 40 epochs with a batch size of 64, using a learning rate of 0.0005 and the AdamW optimizer. We set the temperature parameters $\tau_1 = \tau_2 = 0.07$ and weighting factors $\lambda_1 = \lambda_2 = 0.1$ in all the experiments.

Disease description and prompt preparation: GPT-4 [1] was employed to draft $\mathtt{Prompt}_k$ with the template outlined in Eq. 1, followed by subsequent manual revisions under the supervision of *a clinical professional*. The prompt examples are given in Supplementary Sec. 1. For the PanNuke dataset, a total of 38 $\mathtt{Prompt}$ templates were created to represent benign and malignant tissues across 19 organs, corresponding to 38 unique learnable disease prototypes. For the COVID-x dataset, 3 $\mathtt{Prompt}$ templates and 3 trainable disease prototypes were developed to depict COVID-19, non-COVID pneumonia, and normal cases respectively.

3.2   Method Effectiveness

We demonstrate the effectiveness of our method by comparing it with adapter-based methods (linear probing [19], CLIP-Adapter [5] and Tip-Adapter [29]), prompting-based approaches (CoOp [31], CoCoOp [30] and MaPLe [13]).

PanNuke. When applied to downstream tasks, PLIP shows suboptimal zero-shot performance ($F1{<}0.5$) on certain organs, including the kidney, colon, and pancreas, which were due to their *under representation* in the pretraining dataset as mentioned by the authors [11]. We fine-tuned the PLIP by sampling 5% of the training data in each run and calculated the 10-run mean metric values on the test set to eliminate the randomness. The results are presented in Table 1. The adapter-based methods generally performed the worst among all methods. This indicates that simply tweaking the last layer may not effectively leverage the VLMs' capabilities in the downstream task. Our approach significantly ($p < 0.05$ with Student's t-test) outperformed other prompting-based methods in almost all the cases. This underscores DiCoP's advantage over optimizing prompts without clinical knowledge, and demonstrates DPL's benefit in tailoring VLMs to less represented organs, thereby avoiding overfitting on scarce data.
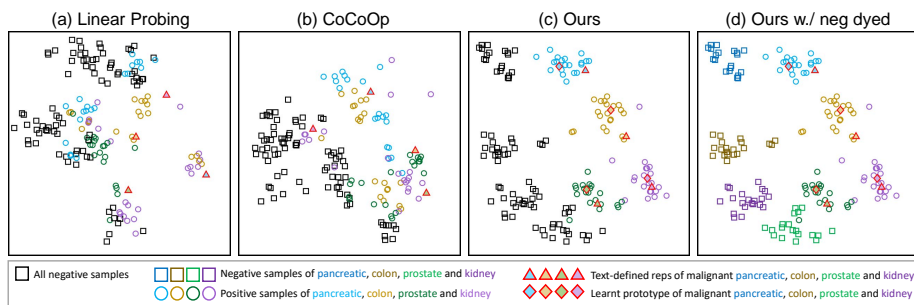
Fig. 2. t-SNE visualization on the latent space on PanNuke. **(a)** Linear probing, **(b)** CoCoOp, **(c)** our method, and **(d)** ours with negative samples of each organ dyed distinctly.

Fig. 2 visualizes the geometric structure of the learned representations of different models via t-SNE [17]. Compared to **(a)** linear probing and **(b)** CoCoOp, our method in **(c)** produced encouraging results in superior separation between positive and negative samples. That explains the excellent performance in Table. 1. In addition, our method presents compact organ-specific clusters, which is attributed to DPL. The clustered representations can be also extended to negative samples by coloring each organ differently as shown in **(d)**. In contrast to our method, CoCoOp does not properly align the positive samples (∘) of each organ with the representations of their corresponding text-only prompt (△). This misalignment indicates CoCoOp's weakness in lacking the clinical knowledge from the textual input and geometric regularizations in the latent space.

Table 2. 10-run average results of COVID finding and all findings on COVID-x with 1% training data. The best performance is in **bold** and the second best is underlined.

Method	COVID-19 finding		All findings	
	Precision	Recall	Acc	Weighted F1
Linear Probing	0.668*	0.601*	0.623*	0.621*
CLIP–Adapter	0.675*	0.645*	0.638*	0.640*
Tip–Adapter	0.682*	0.667*	0.640*	0.641*
CoOp	0.665*	0.628*	0.638*	0.636*
CoCoOp	0.603*	0.703*	0.662*	0.658*
MaPLe†	-	-	-	-
DiCoP + DPL (ours)	**0.702**	**0.770**	**0.741**	**0.744**

* $p < 0.05$ in the one-tailed paired Student's t-test with 10-run results of our method.
† MaPLe cannot directly be applied on the CNN-based VLM.

COVID-x. BioViL [2], pretrained on the MIMIC dataset [12], was developed prior to the COVID pandemic. This makes COVID-19 a *new* disease for the model. Similarly, we did 10-run training with 1% random sampling of the training data. Compared to other methods, Table 2 shows the effectiveness of our method
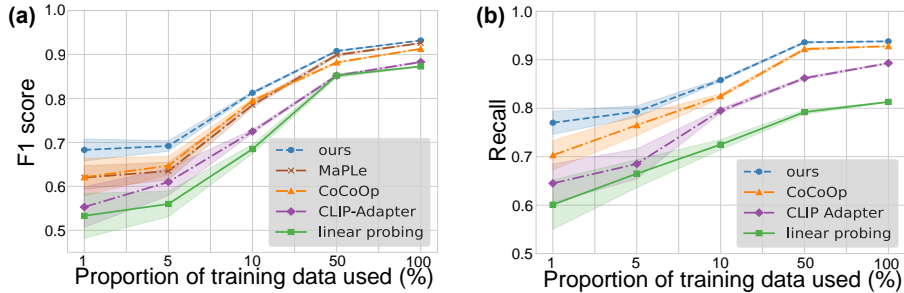
Fig. 3. Data efficiency on **(a)** kidney cancer in PanNuke and **(b)** COVID in COVID-x.
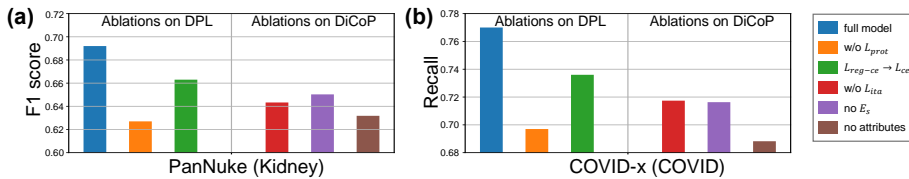

Fig. 4. Ablation studies on **(a)** kidney in PanNuke and **(b)** COVID in COVID-x.

with significantly higher average precision and recall on detecting COVID-19, as well as significantly better accuracy and weighted-F1 scores across all classes. The results underscore the benefits of explaining diseases with clinical language by DiCoP and disease concept learning via DPL.

3.3   Data Efficiency

Considering the data scarcity of *underrepresented* or *newly emerged* diseases, we examine data efficiency by testing the model performance with different proportions of training data. Fig. 3 **(a)** and **(b)** display our method's superior 10-run average F1 score of the kidney tissue classification in PanNuke and recall ratio of the COVID class in COVID-x, respectively.Notably, our approach surpasses other baselines, particularly with limited training data.

3.4   Ablation Study

To evaluate the efficacy of the each component in our method, we conducted an ablation study on the kidney tissue classification in PanNuke (5% training data) and COVID detection in COVID-x (1% training data) As shown in Fig. 4, the study includes: **1)** excluding each loss (w/o L_{prot}, w/o L_{ita}); **2)** replacing L_{reg-ce} with normal CE loss ($L_{reg-ce} \rightarrow L_{ce}$); **3)** no image-specific features (no E_s); **4)** removing the descriptive attributes from Eq. 1 (no attributes). Each study was done in a leave-one-out fashion. Fig. 4 demonstrates a decline in model performance relative to the full model upon removing or altering each component, indicating the efficacy of all components.

4   Conclusion

This paper introduces a novel framework for adapting pretrained VLMs to *underrepresented* or *new* diseases when data is limited. Empirical analyses validate the effectiveness and efficiency of our method across different VLMs and tasks. Our approach facilitates the disease-informed computer-aided diagnosis, potentially broadening its application to other medical image analysis tasks.

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

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 preprint arXiv:2303.08774 (2023) 6
2. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: European conference on computer vision. pp. 1–21. Springer (2022) 1, 5, 7
3. 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) 5
4. Gamper, J., Alemi Koohbanani, N., Benet, K., Khuram, A., Rajpoot, N.: Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In: Digital Pathology: 15th European Congress, ECDP 2019, Warwick, UK, April 10–13, 2019, Proceedings 15. pp. 11–19. Springer (2019) 5
5. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision **132**(2), 581–595 (2024) 2, 6
6. Goel, S., Bansal, H., Bhatia, S., Rossi, R., Vinay, V., Grover, A.: Cyclip: Cyclic contrastive language-image pretraining. Advances in Neural Information Processing Systems **35**, 6704–6719 (2022) 2, 4
7. Hafner, M., Katsantoni, M., Köster, T., Marks, J., Mukherjee, J., Staiger, D., Ule, J., Zavolan, M.: Clip and complementary methods. Nature Reviews Methods Primers **1**(1), 20 (2021) 1
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 5
9. Howard, J., Ruder, S.: Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146 (2018) 5
10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021) 5
11. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023) 1, 5, 6
12. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019) 5, 7
13. 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) 2, 6
14. 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) 3
15. Luan, T., Li, Z., Chen, L., Gong, X., Chen, L., Xu, Y., Yuan, J.: Spectrum auc difference (saucd): Human-aligned 3d shape evaluation. arXiv preprint arXiv:2403.01619 (2024) 1

16. Luan, T., Zhai, Y., Meng, J., Li, Z., Chen, Z., Xu, Y., Yuan, J.: High fidelity 3d hand shape reconstruction via scalable graph frequency decomposition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16795–16804 (2023) 1
17. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008) 7
18. Qin, Z., Yi, H., Lao, Q., Li, K.: Medical image understanding with pretrained vision language models: A comprehensive study. arXiv preprint arXiv:2209.15517 (2022) 1, 3
19. 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) 1, 2, 6
20. Reiß, S., Seibold, C., Freytag, A., Rodner, E., Stiefelhagen, R.: Every annotation counts: Multi-label deep supervision for medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9532–9542 (2021) 1
21. Shen, S., Li, C., Hu, X., Xie, Y., Yang, J., Zhang, P., Gan, Z., Wang, L., Yuan, L., Liu, C., et al.: K-lite: Learning transferable visual models with external knowledge. Advances in Neural Information Processing Systems **35**, 15558–15573 (2022) 1, 3
22. Tsai, E.B., Simpson, S., Lungren, M.P., Hershman, M., Roshkovan, L., Colak, E., Erickson, B.J., Shih, G., Stein, A., Kalpathy-Cramer, J., et al.: The rsna international covid-19 open radiology database (ricord). Radiology **299**(1), E204–E213 (2021) 2
23. Wang, L., Lin, Z.Q., Wong, A.: Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. Scientific reports **10**(1), 19549 (2020) 5
24. Wang, S., Li, C., Wang, R., Liu, Z., Wang, M., Tan, H., Wu, Y., Liu, X., Sun, H., Yang, R., et al.: Annotation-efficient deep learning for automatic medical image segmentation. Nature communications **12**(1), 5915 (2021) 1
25. Zhang, J., Chao, H., Dasegowda, G., Wang, G., Kalra, M.K., Yan, P.: Revisiting the trustworthiness of saliency methods in radiology ai. Radiology: Artificial Intelligence **6**(1), e220221 (2023) 1
26. Zhang, J., Chao, H., Dhurandhar, A., Chen, P.Y., Tajer, A., Xu, Y., Yan, P.: Spectral adversarial mixup for few-shot unsupervised domain adaptation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 728–738. Springer (2023) 1
27. Zhang, J., Chao, H., Dhurandhar, A., Chen, P.Y., Tajer, A., Xu, Y., Yan, P.: When neural networks fail to generalize? a model sensitivity perspective. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 11219–11227 (2023) 1
28. Zhang, J., Chao, H., Yan, P.: Toward adversarial robustness in unlabeled target domains. IEEE Transactions on Image Processing **32**, 1272–1284 (2023) 1
29. Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:2111.03930 (2021) 2, 6
30. 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) 2, 3, 4, 6
31. 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) 2, 6