Global and Local Vision-Language Alignment for Few-Shot Learning and Few-Shot OOD Detection

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Training data in the medical domain is often limited due to privacy concerns and data scarcity. In such few-shot settings, neural network models are prone to overfitting, resulting in poor performance on new in-distribution (ID) data and misclassification of out-of-distribution (OOD) data as learned ID diseases. Existing research treats these two tasks (few-shot learning and few-shot OOD detection) separately, and no prior work has explored a unified approach to simultaneously improving the performance of both tasks. To bridge this gap, we propose a novel framework based on CLIP that jointly enhances ID classification accuracy and OOD detection performance. Our framework consists of three key components: (1) a visually-guided text refinement module, which refines text representations of each disease utilizing disease-relevant visual information; (2) a local version of supervised contrastive learning, which enhances local representation consistency among disease-relevant regions while improving ID-OOD separability; and (3) a global and local image-text alignment strategy, which adaptively combines the global and local similarity measurements for better image-text alignment. Extensive experiments demonstrate that our method outperforms the best methods specifically-tailored for both tasks, achieving new state-of-the-art performance. The source code will be publicly released.

Keywords: Few-shot learning · OOD detection · Disease diagnosis.

1 Introduction

The success of deep learning is often built on large-scale training data [4, 7]. However, in the medical domain, training data is often limited and may not encompass all disease types due to collection difficulties [15]. This not only often leads to model overfitting, but also may result in misclassifying data from unknown diseases as learned diseases, posing serious safety concerns [23, 17]. Various few-shot learning (FSL) techniques have been developed to alleviate the overfitting issue [9, 16, 32], while few-shot out-of-distribution (OOD) detection techniques have been developed to help the model identify whether a new image is from one learned disease or from an unknown disease [5, 19, 27].

In FSL, recent approaches utilize the prior knowledge in pre-trained models to alleviate model overfitting [16, 26, 35]. In particular, pre-trained Vision-Language

Models (VLMs) such as CLIP [24] have demonstrated strong potential in FSL, where parameter-efficient fine-tuning methods (e.g., prompt learning [12, 38, 39], adapters [6, 36], etc.) have been explored to adapt CLIP to downstream few-shot image classification tasks, achieving excellent performance on natural images. In the medical scenario, researchers have also validated the feasibility of adapting CLIP models to medical image classification tasks [25, 37]. However, the performance of adpated CLIP remains unsatisfactory due to the domain gap. On the other hand, models trained on limited data often exhibit poor OOD detection performance due to inaccurate learning of known in-distribution (ID) classes. To tackle this challenge, few-shot OOD detection techniques have been proposed, aiming to improve the model's OOD detection ability in few-shot setting. These methods are primarily based on additional OOD prompts [2, 14, 21] or OOD regularization during model training [20, 34]. However, they struggle to capture the subtle ID-OOD differences among diseases in the medical domain, leading to poor OOD detection performance. While many studies have explored FSL and few-shot OOD detection separately, no prior work has explored a unified approach to improve both ID classification and OOD detection performance, particularly in the medical domain.

In this study, we propose a novel framework to jointly enhance the few-shot image classification and few-shot OOD detection performance. This framework consists of three key components: (1) a visually-guided text refinement module, which refines text representations based on disease-relevant information to improve image-text consistency; (2) a local version of supervised contrastive learning method, which innovatively uses estimated disease relevant and irrelevant regions to enhance local representation consistency within each disease while improving ID-OOD separability; and (3) an global and local image-text alignment strategy, which adaptively adjusts the importance of global and local representations during similarity measurement to achieve better image-text alignment. Extensive experiments on three medical benchmarks demonstrate that our method outperforms the best FSL and few-shot OOD detection methods by a large margin, achieving new state-of-the-art performance in both tasks.

2 Method

Motivation. We propose a novel framework for simultaneous few-shot medical image classification and few-shot OOD detection built on the vision-language model CLIP (Figure 1). Previous CLIP-based methods heavily depend on good image-text alignments, which are often based on the visual embedding of the whole input image and the textual embedding of certain per-class description. However, in the medical domain, the region of interest (ROI, e.g., lesion region) may only occupy a small part of the whole input image, and per-class description may not well capture multiple symptoms appearing in the same disease. With these considerations in mind, we propose utilizing multiple symptom descriptions for each disease, refining the textual embedding from the original CLIP's text encoder (Section 2.1), exploiting local image regions for discriminative lo-

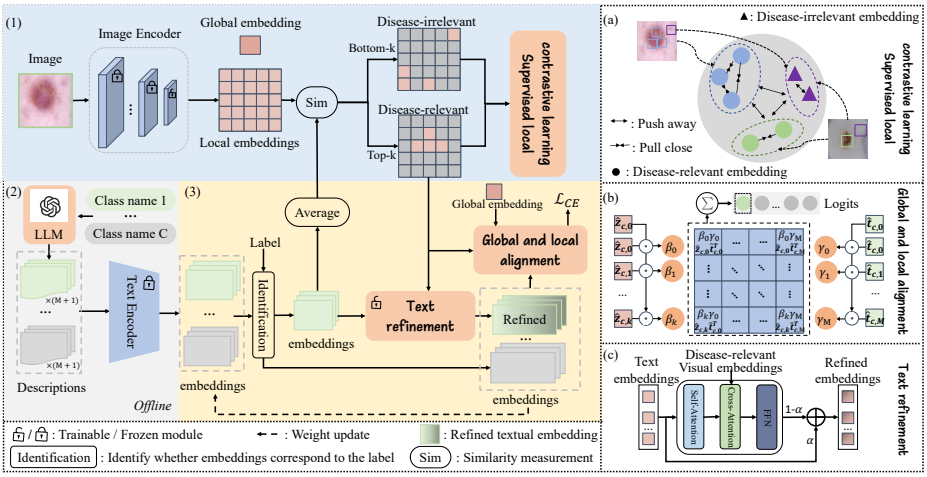

Fig. 1. Overview of our framework: (1) The proposed supervised local contrast learning enhances disease representation consistency and improves ID-OOD separability using disease relevant and irrelevant visual embeddings. (2) LLM generates detailed descriptions for each disease. (3) Text embeddings of descriptions are refined, which, along with disease-relevant visual embeddings are used to improve image-text alignment.

cal representation and better separation between known diseases and unknown diseases (Section 2.2), and adjusting image-text alignment by considering both global and local information in images and texts (Section 2.3). Such modifications are expected to improve performance of the vision-language model in both few-shot medical image classification and few-shot OOD detection.

2.1 Refined textual representations of each disease

Considering that the original CLIP may struggle to understand the semantic information of medical terms (e.g., disease names) simply based on the text input "A photo of {Disease Name}.", we obtain detailed disease descriptions from a large language model (LLM) Qwen-turbo [1] to represent the various symptoms for each disease. Specifically, a short summary ('Dataset Description') of the medical training dataset is first provided which mainly contains information about the number and names of diseases. Then, with the input prompt "Generate questions to classify images which are from {Dataset Description}.", The LLM is asked to generate targeted questions for describing each disease. These generated questions then serve as input for the LLM to generate detailed disease descriptions. For each disease, a total of M descriptions are generated by the LLM, along with "A photo of {Disease Name}", resulting in $M+1$ descriptions.

However, some of the generated descriptions may not accurately describe corresponding disease symptoms or even fail to capture certain disease symptoms that are difficult to describe verbally. Here, a visually-guided text refinement

module (Fig. 1, (c)) is proposed to adjust the textual embedding of each description for each disease based on disease-relevant visual information extracted from training images of the same disease.

Formally, given a training image $\mathbf{x}$ from class c, let $\mathbf{Z} = [\mathbf{z}_0, ..., \mathbf{z}_N] \in \mathbb{R}^{(N+1)\times d}$ and $\mathbf{T}_c = [\mathbf{t}_{c,0}, ..., \mathbf{t}_{c,M}] \in \mathbb{R}^{(M+1)\times d}$ respectively denote the visual embeddings from the CLIP's image encoder and the textual embeddings of the $M+1$ descriptions from the CLIP's text encoder. Here, $\mathbf{z}_0$ is the global image embedding and the other $\mathbf{z}_i$'s are visual embeddings of local image regions. $\mathbf{t}_{c,0}$ is the textual embedding of the default prompt input "A photo of {Disease Name}." and the other $\mathbf{t}_{c,j}$'s correspond to the M descriptions generated from the LLM. d is the embedding dimension for both images and texts. To refine the textual embeddings with disease-relevant visual information from the input image, the irrelevant visual information appearing in the background (i.e., non-lesion) regions should be discarded. Here, a subset of k local visual embeddings which are more likely from lesion regions are simply selected based on certain similarity measurement (cosine similarity by default) between each visual embedding $\mathbf{z}_i$ $(i = 1, \ldots, N)$ and the average textual embedding $\bar{\mathbf{t}}_c = \frac{1}{M+1}\sum_{j=0}^{M} \mathbf{t}_{c,j}$. The selected top-$k$ visual embeddings and the $M+1$ textual embeddings are fed into the proposed 'Text refinement' module (Fig. 1, (c)) which includes a self-attention layer, a cross-attention layer, and two-layer feed-forward network with layer normalization. In particular, for the cross-attention layer, textual embeddings are used as queries and the visual embeddings are used as keys and values. The output $\mathbf{t}'_{c,j}$ from the text refinement module and the corresponding textual embedding $\mathbf{t}c_j$ are fused in the form $\hat{\mathbf{t}}_{c,j} = \alpha\mathbf{t}_{c,j} + (1-\alpha)\mathbf{t}'_{c,j}$ as the final textual embedding for the j-th text description of class c, where α is a weight constant to balance the two terms. Note that the fused textual embeddings $\{\hat{\mathbf{t}}_{c,j}, j = 0, \ldots, M\}$ will replace the original textual embeddings $\{\mathbf{t}_{c,j}, j = 0, \ldots, M\}$ to obtain the average textual embedding $\bar{\mathbf{t}}_c$ for selecting top-k visual embeddings in the next iteration during model training. With the text refinement module, it is expected that the refined textual embeddings $\{\hat{\mathbf{t}}_{c,j}, j = 0, \ldots, M\}$ will be better aligned with the visual representations of disease symptoms, and therefore more helpful in recognizing known diseases and differentiating unknow diseases from known diseases.

2.2 Supervised local contrastive learning

Existing studies often extract the visual representation from the whole image, overlooking the semantic differences among local image regions [11]. In medical images, foreground regions (e.g., lesions) contain critical diagnostic features, while background regions can be viewed as disease-irrelevant OOD signals. Here, we propose a local version of supervised contrastive learning (Fig. 1, (a)), which uses estimated disease-relevant local regions to enhance representation consistency within each class and uses estimated background regions to help improve OOD detection.

Specifically, for each training image $\mathbf{x}$, besides top-k disease-relevant visual embeddings which are selected as described above, another subset of k (bottom-

k) visual embeddings which are most dissimilar to the average textual embedding are selected in a similar way. Given a mini-batch training set, suppose S is the total number of selected top-k and bottom-k visual embeddings from all images of the mini-batch set. Then the supervised local contrastive loss is designed as

$$\mathcal{L}_{SC} = -\frac{1}{S}\sum_{s=1}^{S}\frac{1}{|P(s)|}\sum_{p\in P(s)}\log\frac{\exp\left(g(\mathbf{z}_s,\mathbf{z}_p)/\tau\right)}{\sum_{a=1}^{S}\exp\left(g(\mathbf{z}_s,\mathbf{z}_a)/\tau\right)}, \tag{1}$$

where $P(s)$ is the set of visual embedding indices in which the corresponding visual embeddings share the same class label as that of the visual embedding with index s. The collection of bottom-k visual embeddings from all images in the mini-batch set share a special new class label (i.e., the OOD class). $g(\cdot,\cdot)$ is a similarity measurement function (cosine similarity by default), and τ is the temperature scaling factor. By minimizing $\mathcal{L}_{SC}$, the visual embeddings of lesion regions are more consistent within each disease and different across diseases, which are expected to help improve disease classification and OOD detection.

2.3 Global and local image-text alignment

The global image-text alignment based on the cosine similarity between the visual embedding of the whole image and the textual embedding of a single class description is commonly used for image classification in CLIP-based methods. However, not all image regions contribute equally to classification and some regions may introduce distracting signals [13]. A similar issue may appear when multiple text descriptions are available for each class. Considering such issues, we propose an adaptive global and local image-text alignment strategy (Fig. 1, (b)) to fully utilize the sets of visual embeddings and textual embeddings for better image-text alignment.

Formally, given the set of visual embeddings $\mathbf{Z} = [\mathbf{z}_0, \ldots, \mathbf{z}_N]$ from the image encoder for any input image $\mathbf{x}$ and the set of refined textual embeddings $\hat{\mathbf{T}}_c = [\hat{\mathbf{t}}_{c,0}, \ldots, \hat{\mathbf{t}}_{c,M}]$ of any class c, the top-k visual embeddings $\hat{\mathbf{Z}}_c = \{\hat{\mathbf{z}}_{c,i}, i = 1, \ldots, k\}$ are first selected as described above. The importance for each selected visual embedding and each textual embedding is estimated respectively with

$$\beta_i = \frac{\exp(h(\hat{\mathbf{z}}_{c,i}, \mathbf{z}_0))}{\sum_{a=0}^{k}\exp(h(\hat{\mathbf{z}}_{c,a}, \mathbf{z}_0))}, \ \gamma_j = \frac{\exp(h(\hat{\mathbf{t}}_{c,j}, \hat{\mathbf{t}}_{c,0}))}{\sum_{a=0}^{M}\exp(S(\hat{\mathbf{t}}_{c,a}, \hat{\mathbf{t}}_{c,0}))}, \tag{2}$$

where $h(\cdot,\cdot)$ is certain similarity measurement function (cosine similarity by default). Note that $\mathbf{z}_0$ and $\hat{\mathbf{t}}_{c,0}$ respectively represent the visual embedding of the whole image and the refined textual embedding of the description based on class name. Therefore, higher β_i indicates that $\hat{\mathbf{z}}_{c,i}$ is more important for expressing the semantics of the whole image, and higher γ_j suggest a stronger correlation between $\hat{\mathbf{t}}_{c,j}$ and the class label. The image-text alignment is then defined as

$$u_c = \sum_{i=0}^{k}\sum_{j=0}^{M}\beta_i\gamma_j h(\hat{\mathbf{z}}_{c,i}, \hat{\mathbf{t}}_{c,j}), \tag{3}$$

where the special $\hat{\mathbf{z}}_{c,0}$ is $\mathbf{z}_0$. Such weighted sum of similarity measurements can reduce adverse impact of noisy visual and textual embeddings, which is expected to achieve better image-text alignment for classification ID and OOD detection.

2.4 Model training and Inference

During model training, only the last layer of the image encoder and the text refinement module are trained, while the rest of the model remains frozen. To mitigate overfitting, two $L1$ regularization terms $\mathcal{L}_I$ and $\mathcal{L}_T$ are also adopted following the work [10], where the two terms respectively impose a constraint on the updated visual and text embeddings to ensure their consistency with the corresponding original CLIP's embeddings. Overall, the total loss function for training is $\mathcal{L} = \mathcal{L}_{CE} + \lambda_1\mathcal{L}_{SC} + \lambda_2\mathcal{L}_I + \lambda_3\mathcal{L}_T$, where $\lambda_1, \lambda_2, \lambda_3$ are weight constants to balance the loss terms.

Once the model is well trained, the refined textual embeddings for each class are stored. Given any test image, considering that the class of the image is unknown and therefore top-k selection is unavailable, the whole set of visual embeddings $\mathbf{Z}$ is used to compute the weighted image-text alignment similar to Equation (3). The class with the highest alignment score u_c is chosen as the classification result. For OOD detection, essentially any post-hoc method [8] can be applied. Here, the recently proposed MCM score [18], $Q = \max_c \frac{\exp(u_c/\tau')}{\sum_{a=1}^{C}\exp(u_a/\tau')}$, is simply adopted, where τ' is the temperature scaling factor. Lower Q value indicates that the input image is more likely to be OOD.

3 Experiments

3.1 Experimental Setup

Benchmarks. The proposed method was evaluated on three benchmarks: Skin40, RFMiD6, and ISIC8. Each benchmark consists of an ID training set, an ID test set, and an OOD test set. For the Skin40 [33] benchmark, 40 classes of skin disease images from the SD-198 [28] were used as ID classes and images from the remaining 158 classes of SD-198 were used as the OOD set. For the RFMiD6 benchmark, 6 classes of ocular disease images from the publicly available dataset RFMiD [22] were used as ID classes (branch retinal vein occlusion, diabetic retinopathy, macular hole, myopia, optic disc cupping, optic disc edema). The other 6 classes were used as OOD classes (age-related macular degeneration, central serous retinopathy, disease risk, drusen, optic disc pallor, retinitis). Multi-labeled images were removed. For the ISIC8 benchmark, 8 classes of skin diseases with dermoscopic images were used as ID classes [30], and DermNet [3] was used as the OOD set after the images from ID classes were removed.

Implementation. In all experiments, the pre-trained CLIP model based on ViT-B/16 was used as the backbone. SGD with an initial learning rate of 0.01 was used as the optimizer, and the learning rate was decayed according to the cosine annealing schedule. The batch size was set to 32, and the model was

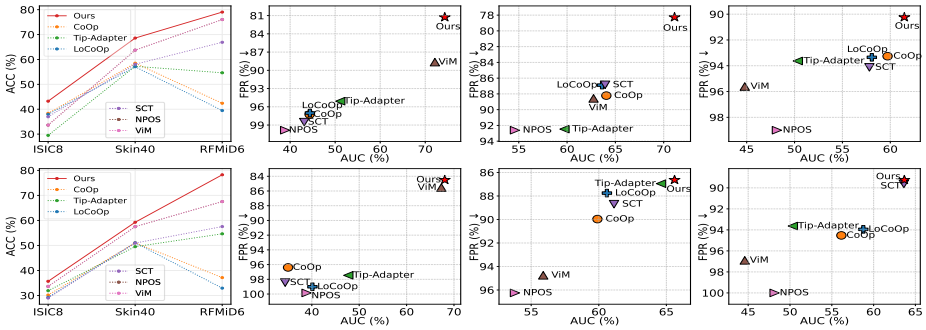

Fig. 2. Performance of few-shot classification (column 1) and OOD detection (columns 2-4) on the ISIC8 (column 2), Skin40 (column 3) and RFMiD6 benchmark (column 4), under 16-shot (row 1) and 8-shot (row 2) settings.

fine-tuned for 200 epochs. For hyperparameter settings, k was set to 50. The weights $\lambda_1, \lambda_2, \lambda_3$ were set to 10, 25, and 1, respectively. The fusion coefficient α was set to 0.99. The temperature scaling factors τ and τ' are set to 1. For few-shot learning, following previous studies [39], a number of 8 and 16 shots were respectively used for training, excluding extreme settings (i.e., 1-, 2-, 4-shot). The evaluation metrics include accuracy (ACC) for classification, and FPR95 (FPR) and AUROC (AUC) for OOD detection. All results were reported as averages over 3 runs with different seeds.

3.2 Result analysis

Efficacy Evaluation. The proposed method was empirically compared with SOTA baselines CoOp [39] and Tip-Adapter [36] for few-shot learning, LoCoOp [20] and SCT [34] for few-shot OOD detection, ViM [31] and NPOS [29] for traditional OOD detection. As shown in Fig. 2, our method consistently outperforms all baselines specifically designed for few-shot classification and OOD detection across all benchmarks under both 16-shot and 8-shot settings, demonstrating its effectiveness.

Ablation of key components. To evaluate the effect of each component in the proposed framework, extensive ablation studies were performed on the representative benchmark (Skin40). As shown in Table 1, when more components are included (columns 1-5), the few-shot classification and few-shot OOD detection performance is gradually improved. In addition, when each component is removed individually (columns 6-8), the performance clearly decreased compared to our full version (column 5). These results demonstrate the effectiveness of the four components in improving both few-shot classification and OOD detection performance.

Effect of global and local image-text alignment. Additional ablation experiments were performed to verify the the necessity of selecting top-k visual embeddings and the need of weights β_i and γ_j (Equation 3). As shown in Fig. 3,

Table 1. Ablation study of key components on Skin40 dataset under the 16-shot setting. 'LLM': inclusion of LLM. 'TRef.': text refinement module; 'LocSC.': supervised local contrastive learning; 'LocAli.': local image-text alignment.

LLM		✓	✓	✓	✓	✓	✓	
TRef.			✓	✓	✓	✓		✓
LocSC.				✓	✓		✓	✓
LocAli.					✓	✓	✓	✓
ACC ↑	$65.80_{\pm 0.83}$	$66.00_{\pm 1.80}$	$66.73_{\pm 0.73}$	$68.33_{\pm 0.90}$	**68.60**$_{\pm 0.91}$	$67.47_{\pm 0.37}$	$66.90_{\pm 0.79}$	$67.40_{\pm 1.27}$
FPR ↓	$87.06_{\pm 1.41}$	$86.32_{\pm 0.54}$	$85.34_{\pm 0.54}$	$83.40_{\pm 0.82}$	**78.26**$_{\pm 1.21}$	$83.11_{\pm 0.82}$	$81.97_{\pm 0.63}$	$89.41_{\pm 1.58}$
AUC ↑	$63.70_{\pm 1.04}$	$64.21_{\pm 0.56}$	$67.15_{\pm 0.56}$	$67.51_{\pm 0.71}$	**71.12**$_{\pm 0.30}$	$70.78_{\pm 0.71}$	$70.88_{\pm 0.23}$	$63.57_{\pm 0.39}$

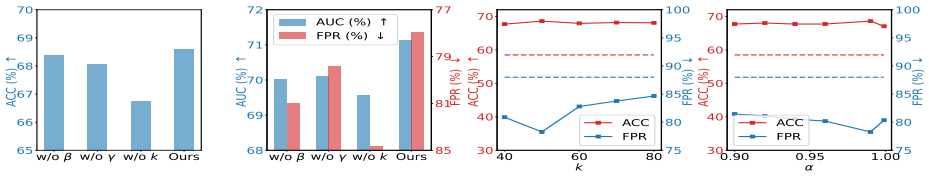

Fig. 3. Left half: effect of top-k selection and weights in alignment for classification and OOD detection. 'w/o β' and 'w/o γ': all β_i's and γ_j's in Equation (3) are fixed to 1.0; 'w/o k': all image regions are used. Right half: sensitivity study of k in top-k selection and α for text refinement.

the ablated versions performed worse than our full version, confirming the effectiveness of these components. Notably, when top-k embeddings were not selected, the performance decreased significantly, confirming the importance of excluding disease-irrelevant regions when measuring the image-text similarity.

Sensitivity studies. To evaluate the insensitivity of value choices for the hyper-parameters k in top-k selection and α in text refinement, a series of sensitive studies were performed on Skin40. As shown in Fig. 3, when k varies in the range [20, 80] and α varies in the range [0.88, 1), the performance of our method (solid curves) remains relatively stable and surpasses the strong baseline CoOp (dashed lines), suggesting that our method is insensitive to hyper-parameter choice.

4 Conclusion

In this study, a novel CLIP-based training and inference framework is proposed for few-shot medical image classification and few-shot OOD detection. The text refinement module refines text embeddings using estimated disease-relevant visual information. The supervised local contrastive learning uses estimated disease-relevant and background local regions to enhance representation consistency within each class and to help improve OOD detection. Additionally, the proposed global and local image-text alignment achieves better alignment of local visual-textual sets, further improving ID classification and OOD detection. Our method achieves new SOTA performance in both tasks on three medical benchmarks and holds promise for extension to more imaging modalities.

References

1. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
2. Bai, Y., Han, Z., Cao, B., Jiang, X., Hu, Q., Zhang, C.: Id-like prompt learning for few-shot out-of-distribution detection. In: CVPR (2024)
3. Bajwa, M.N., Muta, K., Malik, M.I., Siddiqui, S.A., Braun, S.A., Homey, B., Dengel, A., Ahmed, S.: Computer-aided diagnosis of skin diseases using deep neural networks. Applied Sciences **10**(7), 2488 (2020)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
5. Chen, X., Li, Y., Chen, H.: Dual-adapter: Training-free dual adaptation for few-shot out-of-distribution detection. arXiv preprint arXiv:2405.16146 (2024)
6. 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)
7. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
8. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: ICLR (2017)
9. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: CVPR (2023)
10. Khattak, M.U., Wasim, S.T., Naseer, M., Khan, S., Yang, M.H., Khan, F.S.: Self-regulating prompts: Foundational model adaptation without forgetting. In: CVPR (2023)
11. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. NeurIPS (2020)
12. Lafon, M., Ramzi, E., Rambour, C., Audebert, N., Thome, N.: Gallop: Learning global and local prompts for vision-language models. In: ECCV (2024)
13. Li, J., Li, H., Erfani, S.M., Feng, L., Bailey, J., Liu, F.: Visual-text cross alignment: Refining the similarity score in vision-language models. In: ICML (2024)
14. Li, T., Pang, G., Bai, X., Miao, W., Zheng, J.: Learning transferable negative prompts for out-of-distribution detection. In: CVPR (2024)
15. Lin, Y., Chen, Y., Cheng, K.T., Chen, H.: Few shot medical image segmentation with cross attention transformer. In: MICCAI (2023)
16. Liu, F., Zhang, T., Dai, W., Zhang, C., Cai, W., Zhou, X., Chen, D.: Few-shot adaptation of multi-modal foundation models: A survey. Artificial Intelligence Review **57**(10), 268 (2024)
17. Marimont, S.N., Siomos, V., Tarroni, G.: Mim-ood: Generative masked image modelling for out-of-distribution detection in medical images. In: MICCAI (2023)
18. Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., Li, Y.: Delving into out-of-distribution detection with vision-language representations. NeurIPS (2022)
19. Miyai, A., Yang, J., Zhang, J., Ming, Y., Lin, Y., Yu, Q., Irie, G., Joty, S., Li, Y., Li, H., et al.: Generalized out-of-distribution detection and beyond in vision language model era: A survey. arXiv preprint arXiv:2407.21794 (2024)
20. Miyai, A., Yu, Q., Irie, G., Aizawa, K.: Locoop: Few-shot out-of-distribution detection via prompt learning. NeurIPS (2023)
21. Nie, J., Zhang, Y., Fang, Z., Liu, T., Han, B., Tian, X.: Out-of-distribution detection with negative prompts. In: ICLR (2024)

22. Pachade, S., Porwal, P., Thulkar, D., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Giancardo, L., Quellec, G., Mériaudeau, F.: Retinal fundus multi-disease image dataset (rfmid): A dataset for multi-disease detection research. Data **6**(2), 14 (2021)
23. Parnami, A., Lee, M.: Learning from few examples: A summary of approaches to few-shot learning. arXiv preprint arXiv:2203.04291 (2022)
24. 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: ICML (2021)
25. Shakeri, F., Huang, Y., Silva-Rodríguez, J., Bahig, H., Tang, A., Dolz, J., Ben Ayed, I.: Few-shot adaptation of medical vision-language models. In: MICCAI (2024)
26. Song, Y., Wang, T., Cai, P., Mondal, S.K., Sahoo, J.P.: A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. ACM Computing Surveys **55**(13s), 1–40 (2023)
27. Sun, H., He, R., Han, Z., Lin, Z., Gong, Y., Yin, Y.: Clip-driven outliers synthesis for few-shot ood detection. arXiv preprint arXiv:2404.00323 (2024)
28. Sun, X., Yang, J., Sun, M., Wang, K.: A benchmark for automatic visual classification of clinical skin disease images. In: ECCV. pp. 206–222. Springer (2016)
29. Tao, L., Du, X., Zhu, X., Li, Y.: Non-parametric outlier synthesis. arXiv preprint arXiv:2303.02966 (2023)
30. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018)
31. Wang, H., Li, Z., Feng, L., Zhang, W.: Vim: Out-of-distribution with virtual-logit matching. In: CVPR (2022)
32. Wang, Y., Yao, Q., Kwok, J.T., Ni, L.M.: Generalizing from a few examples: A survey on few-shot learning. ACM computing surveys (csur) **53**(3), 1–34 (2020)
33. Yang, Y., Cui, Z., Xu, J., Zhong, C., Zheng, W.S., Wang, R.: Continual learning with bayesian model based on a fixed pre-trained feature extractor. Visual Intelligence **1**(1), 5 (2023)
34. Yu, G., Zhu, J., Yao, J., Han, B.: Self-calibrated tuning of vision-language models for out-of-distribution detection. NeurIPS (2025)
35. Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
36. Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free adaption of CLIP for few-shot classification. In: ECCV (2022)
37. 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)
38. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: CVPR (2022)
39. 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)