

Vision-Language Dual-Pattern Matching for Out-of-Distribution Detection

Zihan Zhang^{*} , Zhuo Xu^{*} , and Xiang Xiang^{*(✉)} 

National Key Lab of Multi-Spectral Information Intelligent Processing Technology,
School of Artificial Intelligence and Automation,
Huazhong University of Science and Technology, Wuhan, China
`xex@hust.edu.cn`

Abstract. Out-of-distribution (OOD) detection is a significant challenge in deploying pattern recognition and machine learning models, as models often fail on data from novel distributions. Recent vision-language models (VLMs) such as CLIP have shown promise in OOD detection through their generalizable multimodal representations. Existing CLIP-based OOD detection methods only utilize a single modality of in-distribution (ID) information (*e.g.*, textual cues). However, we find that the ID visual information helps to leverage CLIP’s full potential for OOD detection. In this paper, we pursue a different approach and explore the regime to leverage both the visual and textual ID information. Specifically, we propose Dual-Pattern Matching (DPM), efficiently adapting CLIP for OOD detection by leveraging both textual and visual ID patterns. DPM stores ID class-wise text features as the textual pattern and the aggregated ID visual information as the visual pattern. At test time, the similarity to both patterns is computed to detect OOD inputs. We further extend DPM with lightweight adaptation for enhanced OOD detection. Experiments demonstrate DPM’s advantages, outperforming existing methods on common benchmarks. The dual-pattern approach provides a simple yet effective way to exploit multi-modality for OOD detection with vision-language representations.

Keywords: Out-of-distribution detection · Vision-language models

1 Introduction

In practical applications of pattern recognition and machine learning, a pervasive assumption is that all encountered classes are in-distribution (ID). However, this assumption is often invalid in real-world scenarios [19, 41, 44, 45], where out-of-distribution (OOD) data, or data belonging to classes not encountered during training, are present. Ensuring the reliability and robustness of model predictions in such environments is a significant challenge [15], as models often produce over-confident, incorrect predictions when faced with OOD data [30].

^{*} Equal contribution, co-first author; Xiang Xiang is also with Peng Cheng Laboratory.

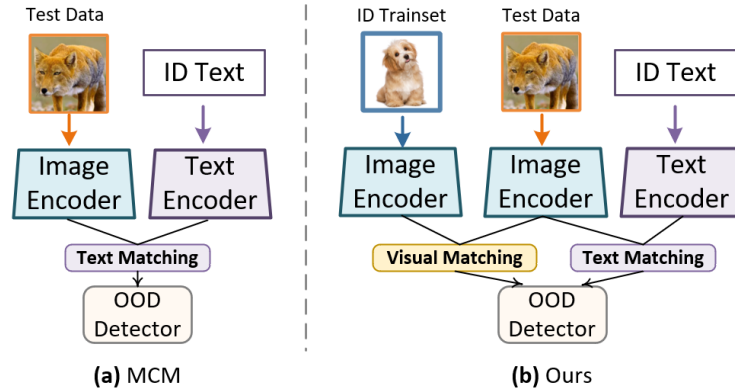


Fig. 1: Comparisons of different paradigms of CLIP-based OOD detection methods.

With the emergence of models trained on extensive datasets, many downstream applications benefit from the transferable representations of the models [22, 32, 48, 57]. Therefore, adapting the pre-trained models for transfer learning is becoming common, rather than training from scratch. In computer vision, the Vision-Language Models (VLMs) show impressive generalizability and zero-shot capabilities across tasks. [32, 48]. Compared to the image pre-training models, VLMs trained on image-text pair data can span a broad range of concepts in the real world, due to the unconstrained nature of textual descriptions.

However, zero-shot recognition typically necessitates the availability of label names for the test data. This is problematic in the context of OOD detection tasks because the OOD label space can be vast. Enumerating the entirety of the OOD space would substantially escalate both storage and inference demands, rendering such an approach impractical. Thus, it becomes crucial to devise dedicated OOD detection algorithms for VLMs to address this challenge.

The evolving field of OOD detection with vision-language representations can be roughly divided into two parts: training-required and training-free methods. Esmailpour *et al.* [10] introduce ZOC, which harnesses the visual encoder of the CLIP model [32] to learn representations of OOD class names. From the view to enabling CLIP to say no, Wang *et al.* [43] present CLIPN, a method that employs negation-semantic prompts to identify OOD samples. Ming *et al.* [27] further delved into the possibilities of parameter-efficient fine-tuning of the CLIP model, demonstrating that fine-tuning could enhance performance on both in-distribution (ID) and OOD tasks. Different from the training-required methods, MCM [26] explores the use of Softmax scaling methods for training-free OOD detection, utilizing the pre-trained CLIP model. These methods use the text features as the cosine classifier to calculate the prediction logits. The OOD detection task can be formulated as a "store and compare" process [51]. A key issue is to calculate and store the ID patterns, which is used to compare with the test samples. Previous CLIP-based methods overlook the multimodal attribute of

VLMs. They only take into account the similarity between the model’s image outputs and the ID text pattern, while disregarding the ID visual pattern. We then ask the question, can we integrate the metrics from unimodality, namely, not only matching the ID text pattern but also the ID visual pattern?

To this end, we propose Dual-Pattern Matching (DPM), which efficiently adapts CLIP for OOD detection by exploiting multimodal ID data information as shown in Fig. 1. DPM achieve superior performance by aligning both visual and text patterns, and it easily integrates with lightweight adaptation training. We observe that the CLIP feature space of the downstream datasets is narrow, where the distances between different ID classes are small, which makes low ID-OOD separability. Motivated by [12], we propose a training-free feature enhancement module to enhance the ID-class features separability, which is verified important for OOD detection [28, 53]. The proposed DPM calculates the visual pattern and text pattern for ID data. Specifically speaking, we use the text features of the class names with templates as the text pattern. Also, we calculate the visual pattern by accumulating the image-text similarity of ID data. The dual-pattern scheme can utilize the visual and text information of the ID data simultaneously. Meanwhile, the use of visual information could better handle the covariate shift [47], as the text pattern matching could not afford the clue of covariate shift. To further embody the core of our method, we extend DPM to DPM-T, which combines lightweight adaptation with prompt learning. We design a learnable prompt for each class and use the projected image feature after the training-free enhancement to align with the text feature. The introduction of lightweight adaptation can concurrently improve the quality of the dual patterns.

We summarize our contributions as follows.

- We propose a domain-specific features aggregation module to refine the CLIP visual feature for better alignment with the textual features.
- We propose Dual-Pattern (visual and textual) Matching to leverage the full potential of CLIP on OOD detection task. On top of that, we introduce two variants of methods, a training-free DPM-F, and a training-required DPM-T.
- Our DPM-F and DPM-T exhibit state-of-the-art performance on OOD detection benchmarks, demonstrating the superiority of our proposed approaches.

2 Related work

Unimodal OOD Detection. Unimodal OOD detection represents the method that directly adopts the image data only. It can be roughly divided into two categories: outlier-free methods and outlier-exposure [3, 8, 31, 50] methods. Outlier-exposure methods necessitate additional OOD data as outliers to aid the model in discerning between ID and OOD discrepancies. In this paper, we focus on the outlier-free methods which do not rely on any unknown data.

A major branch of outlier-free methods is classification-based methods [47] including confidence enhancement methods [1, 9, 37, 38], and post-hoc detection [13, 15–17, 23, 25, 35, 42, 53]. Post-hoc detection methods can be readily utilized with an already trained model, including MSP [15] that uses the highest

softmax probability as the OOD score, and MaxLogit [13] that utilizes the maximum logit value of the samples. Other examples include ODIN [23] which finds substantial temperature scaling and input disturbance beneficial in distinguishing OOD samples, and Energy [25] which suggests an energy score based on the $\log\text{sumexp}$ of logits for OOD detection.

Cross-modal OOD Detection. Recent vision-language models like CLIP [32] have shown exceptional zero-shot performance. The OOD detection with vision-language representations aims to detect OOD samples that do not belong to the ID classes of interest for a downstream task, instead of the classes used during pre-training [27]. ZOC [10] proposes to learn the representations of OOD class names based on the visual encoder of CLIP. MCM [26] investigates training-free OOD detection with softmax scaling methods based on the pre-trained CLIP model. CLIPN [43] proposes to learn the negation-semantic prompts to identify OOD samples. Ming *et al.* [27] investigate parameter-efficient fine-tuning CLIP model, and they find the fine-tuning could improve both ID and OOD performance. However, the existing methods do not leverage the ID visual information, and the efficient parameter tuning methods are not well-designed for OOD tasks.

Vision-Language Models. Visual-language models such as CLIP [32], trained on extensive image-text data, demonstrate significant promise in the learning of general representations [48]. These transferable representations vastly improve performance on a range of downstream tasks. However, downstream datasets often have a limited number of samples. Therefore, this has led to increased interest in developing strategies for adapting CLIP to downstream tasks. Broadly speaking, such adaptation strategies fall into two categories: strategies that necessitate the training of an adapting module, and strategies that bypass such a need. CoOp [56], for instance, proposes to tune prompts for downstream classification tasks for better performance. DualCoOp [34] presents a framework with dual learnable prompts to efficiently transfer the CLIP model to address multi-label image recognition with limited annotations. Gao *et al.* [11] propose to conduct fine-tuning with feature adapters on either visual or language branches, outperforming context optimization while maintaining a simple design. While tuning CLIP can achieve better performance in downstream tasks, such a process still needs extra training and computational resources. Therefore, Zhang *et al.* [52] propose a training-free CLIP Adapter for better vision-language modeling.

3 Methodology

3.1 Preliminary

CLIP-based classification. CLIP consists of two encoders: visual encoder Φ^V and text encoder Φ^T . The encoders enable CLIP to handle information from both visual and textual modalities. For a downstream task that involves K categories, given a label $y_k \in \mathcal{Y}$, text features are extracted as $F_k^T \in \mathbb{R}^{1 \times C}$. For an image $\mathbf{x} \in \mathcal{X}$ to be recognized, the image features are encoded as $f^V = \Phi^V(\mathbf{x}) \in \mathbb{R}^{H \times W \times C}$ and pooled into its global image feature $F^V \in \mathbb{R}^{1 \times C}$:

$$F_k^T = \Phi^T(\text{prompt}(y_k)), F^V = \text{Pooling}(f^V), \quad (1)$$

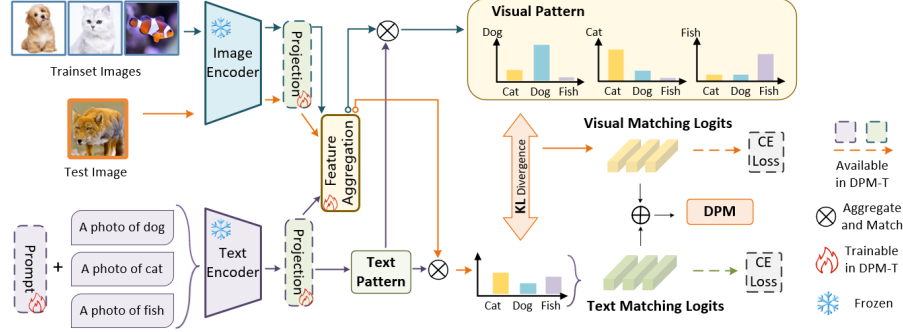


Fig. 2: The pipeline of DPM method. The dashed boxes and dashed arrows indicate the module is only available in the DPM-T. For DPM-F, we add a visual matching brand compared with previous methods. Compared with DPM-T, we add lightweight learnable prompts and a linear projection module after the image encoder.

where $prompt(\cdot)$ is the template for name of input candidate category y_k , for example, {a photo of [Classname]}. Then, the CLIP model outputs the zero-shot classification results based on the cosine similarity between image features and text features with a scaling factor τ .

$$p(y_k | \mathbf{x}) = \frac{\exp(\cos(F^V, F_k^T) / \tau)}{\sum_{j=1}^K \exp(\cos(F^V, F_j^T) / \tau)}. \quad (2)$$

CLIP-based OOD detection. Compared with the unimodal OOD detection, the previous CLIP-based methods treat the text features of each category as the cosine classifier to output the logits or softmax probabilities. Given the outputs, it is easy to apply the post-hoc OOD detection methods such as MSP and Energy.

3.2 Refining CLIP for downstream tasks

For the downstream task, the extracted CLIP visual representations contain much redundant information. For instance, the representations could contain the information of the background, while the foreground information is domain-specific and effective. To this end, we propose a refinement method of Domain-Specific Feature Aggregation to enhance the domain-specific representations.

Domain-specific feature aggregation: The foreground objects appear in different regions of the image, simple pooling methods mix the foreground and background information. The foreground information is domain-specific and effective for classification, therefore, we propose to aggregate the region features extracted by CLIP to get the domain-specific features. As the feature F^V is aligned with the text feature, each region of the image feature map f^V also aligns the text feature [55]. We use $f_r^V \in \mathbb{R}^{1 \times C}$ to denote the feature at spatial location r after f^V is flattened.

To obtain the domain-specific features, we first compute the similarity between the text features F_k^T for each category k and the visual feature of the region r as $\cos(F_r^V, F_k^T)$. Then we apply the similarity on F_r^V to get the features:

$$\bar{F}^V = \sum_{r=1}^{H \cdot W} \frac{\exp(\cos(F_r^V, F_k^T))}{\sum_{j=1}^{H \cdot W} \exp(\cos(F_j^V, F_k^T))} F_r^V \in \mathbb{R}^{1 \times C}. \quad (3)$$

Now the aggregated domain-specific feature is:

$$\tilde{F}^V = \gamma \bar{F}^V + F^V \in \mathbb{R}^{1 \times C}, \quad (4)$$

where γ is a parameter that controls the mixture of the two features. The final classification results are computed as $\tilde{p}(y_k | \mathbf{x}) = \frac{\exp(\cos(\tilde{F}^V, F_k^T)/\tau)}{\sum_{j=1}^K \exp(\cos(\tilde{F}^V, F_j^T)/\tau)}$. Note that domain-specific feature aggregation does not introduce new parameters.

3.3 Dual-Pattern Matching

Previous CLIP-based methods try to use the text pattern as the cosine classifier to calculate the prediction logits. We find that the ID trainset images could serve as a visual pattern in the "store and compare" process [51]. Thus, we propose a new CLIP-based OOD detection method named Dual-Pattern Matching (DPM).

Defining the dual-pattern: We first calculate the visual pattern by accumulating the similarity of the image-text matching score of all ID data. The previous method KNN [36] utilizes the trainset features as the visual pattern for each class. Different from KNN, we use the trainset image-text matching results as the visual patterns for each category because the modality gap exists in the vision-language models [24]. Then, we calculate the similarity between the test image features and visual patterns, combining the results with the textual matching results to get the final results.

Specifically speaking, for a task with K in-distribution classes, We sample n data for each ID category. The text pattern for category k is the text feature $P_k^T = F_k^T = \Phi^T(\text{prompt}(y_k)) \in \mathbb{R}^{1 \times C}$. Given the j_{th} sample of category k as $\mathbf{x}_j$, the image-text matching score is

$$s_j = \text{Concat}[\tilde{p}(y_1 | \mathbf{x}_j), \tilde{p}(y_2 | \mathbf{x}_j), \dots, \tilde{p}(y_K | \mathbf{x}_j)], \quad (5)$$

where $s_j \in \mathbb{R}^{1 \times K}$. The mean of the sampled data's image-text matching score for category k is taken as the visual pattern P_k^V for category k as

$$P_k^V = \frac{1}{n} \sum_{j=1}^n s_j \in \mathbb{R}^{1 \times K}. \quad (6)$$

Now we have a visual pattern P^V and a text pattern P^T for the ID dataset as

$$\begin{aligned} P^V &= \text{Concat}([P_1^V, P_2^V, \dots, P_K^V]) \in \mathbb{R}^{K \times K} \\ P^T &= \text{Concat}([P_1^T, P_2^T, \dots, P_K^T]) \in \mathbb{R}^{K \times C}. \end{aligned} \quad (7)$$

Matching the test feature with dual-pattern: Given an input $\mathbf{x}$ to be recognized, we calculate the two matching scores between the aggregated domain-specific feature $\tilde{F}^V$ of $\mathbf{x}$ and text pattern and visual pattern respectively:

$$s^T = \tilde{F}^V P^{T\top} \in \mathbb{R}^{1 \times K}, s_k^V = \text{KL}(\text{Softmax}(s^T) || P_k^V) \in \mathbb{R}, \quad (8)$$

and $s^V = \text{Concat}([s_1^V, s_2^V, \dots, s_K^V]) \in \mathbb{R}^{1 \times K}$. We use Kullback-Leibler (KL) divergence to measure the similarity between the test features and visual pattern as $\text{KL}(R||Q) = \sum_i R_i \log \frac{R_i}{Q_i}$.

Each pattern matching emphasizes and contributes information from different modalities, therefore, we utilize the two matching scores to help fuse the information from the two modalities. Formally, our OOD detection scoring function DPM can be formulated as:

$$\begin{aligned} DPM(\mathbf{x}) &= \max(s^T) - \beta \min(s^V) \in \mathbb{R}, \\ OOD &= \begin{cases} 1, & \text{if } DPM(\mathbf{x}) > \lambda \\ 0, & \text{if } DPM(\mathbf{x}) < \lambda, \end{cases} \end{aligned} \quad (9)$$

where β is the visual modality affinity factor for the flexibility to balance visual and text information, 1 represents the positive class (ID) and 0 indicates OOD. λ is chosen so that a high fraction of ID data (*e.g.*, 95 %) is above the threshold.

Training-Free DPM-F: Here, we introduce DPM-F, a training-free DPM approach. Our framework is depicted in Fig. 2, where dashed components (learnable prompts and the projection layer) are not included in DPM-F. Importantly, DPM-F doesn't introduce any new trainable parameters. Generally speaking, we add a visual matching branch. The textual matching branch shares a similar architecture with CLIP-based OOD detection methods. Taking the text pattern as a cosine classifier, the branch outputs textual matching logits which can be used for logit-based OOD detection methods. For the visual matching branch, we first aggregate textual matching scores from the trainset to compute the visual pattern for each ID category. Then we measure the similarity between the test image feature and visual pattern to get the visual matching score.

Training-Required DPM-T: To further improve the performance of OOD detection, we introduce a new training-required framework, DPM-T. The domain-specific feature aggregation module helps the visual feature more align with the text pattern, which benefits OOD detection. However, the original text pattern is still unchanged and the distinction between different class text patterns is low.

Introducing learnable parameters: We introduce two learnable modules for the adaptation of both the visual and textual features. Specifically speaking, we add a linear projection layer after the visual encoder for both the visual feature of the region F_r^V and global image feature F^V as $L_s(F_r^V)$, $L_v(F^V)$. The L is an MLP-like projection layer that is appended after the frozen image features to facilitate the matching of pair-wise projected features containing two linear layers, two norm layers and one activation function. For the text features, we add learnable prompts inspired by [56] as $Prompt = [V_1, V_2, V_3, \dots, V_q]$, where $V_1, \dots, V_q$ are q learnable word embedding vectors. Different from using handcraft prompt

templates in DPM-F, we replace them with the learnable Prompt. Therefore, the text input of DPM-T is formed as $[Prompt, Classname]$. To better facilitate the matching of text-vision projected features, we also add a linear layer as $L_t(F^T)$ to get the text pattern. Notably, the parameter number of the linear layer and learnable word embedding vectors is neglectable for the CLIP pre-trained model.

Optimizing the learnable module: One intuitive idea to optimize the module is to match the projected global image feature and the projected text features. However, such optimization does not take into account the process of dual-pattern matching, which results in a performance degradation of the visual-matching branch. Therefore, we directly conduct the textual matching based on Eqs. (3) and (4) and set γ as a learnable parameter. According to the two equations, we replace the features F_r^V, F_k^T and F^V with the projected one and get the text-guided domain-specific feature for class k as $\tilde{F}^{Vt} = L_v(F^V) + \gamma \sum_{r=1}^{H \cdot W} \frac{\exp(\cos(L_s(F_r^V), L_t(F_k^T)))}{\sum_{j=1}^{H \cdot W} \exp(\cos(L_s(F_j^V), L_t(F_k^T)))} L_s(F_r^V) \in \mathbb{R}^{1 \times C}$.

Then, we match the feature with the projected text pattern. For the visual matching branch, we leverage visual prototype matching [28] to encourage train-set data to be close to their class prototypes. The feature compactness in each class benefits the OOD detection, especially for the logit-based methods [53]. Define the visual prototype μ_k of class k as the mean of $\tilde{L}_v(F^V)$ belonging to the class in the trainset. We accumulate the features in the forward process and evaluate the prototype in an exponential moving average (EMA) manner. Without EMA to estimate the prototypes, the prototypes may exhibit fluctuations across different training steps, leading to training instability. Similarly to the $\tilde{F}^{Vt}$, we obtain the vision-guided domain-specific feature for class k as $\tilde{F}^{Vv} = L_v(F^V) + \gamma \sum_{r=1}^{H \cdot W} \frac{\exp(\cos(L_s(F_r^V), \mu_k))}{\sum_{j=1}^{H \cdot W} \exp(\cos(L_s(F_j^V), \mu_k))} L_s(F_r^V) \in \mathbb{R}^{1 \times C}$.

Given the above feature, we conduct the visual and textual matching as

$$p^V(y_k | x) = \underbrace{\frac{\exp(\cos(\tilde{F}^{Vv}, \mu_k) / \tau)}{\sum_{j=1}^K \exp(\cos(\tilde{F}^{Vv}, \mu_j) / \tau)}}_{\text{Visual Matching(VM)}} \quad (10)$$

$$p^T(y_k | x) = \underbrace{\frac{\exp(\cos(\tilde{F}^{Vt}, L_t(F^T)) / \tau)}{\sum_{j=1}^K \exp(\cos(\tilde{F}^{Vt}, L_t(F^T)) / \tau)}}_{\text{Textual Matching(TM)}}.$$

Note that we calculate the cosine similarity between the vision-guided domain-specific feature for class k and μ_k , similar to the text-guided feature. Then, we jointly optimize two losses to promote ID-OOD separability as

$$\mathcal{L} = \mathcal{L}_{VM} + \mathcal{L}_{TM} = \ell_{ce}(p^V, y) + \ell_{ce}(p^T, y). \quad (11)$$

4 Experiments

4.1 Experimental Setup

Out-of-distribution datasets. Following the prior work [43], we use three benchmarks as OOD datasets for CIFAR-100 [21] including CIFAR-10, ImageNet-R [20] and LSUN [49]. For large-scale dataset ImageNet-1k [6], we use iNaturalist [39], SUN [46], Texture [5] and Places365 [54] as OOD datasets to benchmark the algorithms. These OOD datasets provide a comprehensive benchmark for our algorithm and span a diverse range of domains.

Evaluation metrics. We evaluate our methods using two widely accepted metrics: Area Under the Receiver Operating Characteristic Curve (AUROC) and False Positive Rate at 95% True Positive Rate (FPR95). Higher AUROC and lower FPR95 values are desirable. We present both metrics as percentages.

Implementation. All the pre-trained CLIP models utilized in our experiments are obtained from OpenCLIP [4]. Specifically, we use the CLIP models with ViT-B-16 and ViT-B-32 architectures for our evaluations. See also the codes*.

Table 1: Results on CIFAR-100, we search the scaling factor for MCM [26] for each datasets. AUR: AUROC, FPR: FPR95.

Method	CIFAR-10		ImageNet-R		LSUN		Avg.	
	AUR $\uparrow$	FPR $\downarrow$	AUR $\uparrow$	FPR $\downarrow$	AUR $\uparrow$	FPR $\downarrow$	AUR $\uparrow$	FPR $\downarrow$
Image Encoder: ViT-B-16								
MSP [15]	74.53	84.43	58.46	95.59	86.43	62.01	73.14	80.67
MaxLogit [13]	74.36	84.78	84.22	53.33	83.28	85.01	80.62	74.37
Energy [25]	70.88	87.34	84.17	55.07	80.46	89.59	78.50	77.33
ReAct [35]	69.50	93.26	42.62	95.73	84.16	80.91	65.42	89.96
MCM [26]	76.08	85.55	56.81	97.43	88.21	59.02	73.71	80.66
DPM-F	84.81	61.48	84.69	50.95	86.47	59.62	85.32	57.35
Requires training (or w/ fine-tuning)								
CLIPN-A [43]	80.53	69.46	63.79	75.06	89.62	55.83	77.98	66.78
DPM-T	90.55	45.60	90.41	40.28	93.92	30.73	91.62	38.87
Image Encoder: ViT-B-32								
MSP [15]	70.95	87.29	58.14	97.23	88.43	60.63	72.51	81.72
MaxLogit [13]	74.94	87.38	72.51	71.62	90.98	59.71	79.48	78.91
Energy [25]	71.22	86.86	74.78	67.11	85.90	77.99	77.3	79.25
ReAct [35]	68.96	92.45	68.47	75.56	89.47	69.74	75.63	79.25
MCM [26]	72.37	87.86	56.68	98.02	90.57	53.81	73.21	79.89
DPM-F	78.72	74.53	78.30	64.44	95.41	23.81	84.14	54.26
Requires training (or w/ fine-tuning)								
CLIPN-A [43]	88.06	47.99	87.09	60.07	93.55	35.19	89.57	47.75
GROOD [40]	90.72	43.58	82.02	57.25	86.34	65.98	86.36	55.61
DPM-T	94.11	30.80	93.87	28.54	96.39	19.89	94.79	26.41

* <https://github.com/HAIV-Lab/DPM>

Table 2: Results on ImageNet with different backbones. For ReAct, DPM-T, and DPM-F, we only sample 10% of the trainset data for computation and optimization. Specifically, we sample 100 samples for each class in trainset data.

Method	iNaturalist		SUN		Places		Texture		Avg.	
	AUR $\uparrow$	FPR $\downarrow$	AUR $\uparrow$	FPR $\downarrow$	AUR $\uparrow$	FPR $\downarrow$	AUR $\uparrow$	FPR $\downarrow$	AUR $\uparrow$	FPR $\downarrow$
Image Encoder: ViT-B-16										
MSP	83.75	59.18	81.93	61.10	81.11	63.17	79.04	63.14	79.05	63.14
MaxLogit	88.03	60.88	81.32	61.10	80.11	63.17	79.05	64.14	81.06	61.90
Energy	87.54	64.36	87.21	67.26	87.80	56.55	74.89	96.33	84.36	71.13
MCM	90.89	46.51	91.12	42.14	88.68	46.49	87.19	52.38	89.47	46.88
ReAct	89.95	60.45	92.39	43.80	92.01	38.84	93.90	31.05	92.06	43.53
DPM-F	96.94	12.89	92.62	31.63	89.97	41.15	91.60	32.71	92.78	29.59
Requires training (or w/ fine-tuning)										
CLIPN-A	95.27	23.94	93.93	26.17	90.93	40.83	92.28	33.45	93.10	31.10
DPM-T	99.03	5.08	97.07	16.77	91.81	40.54	94.96	21.98	95.72	21.09
Image Encoder: ViT-B-32										
MSP	81.80	65.60	79.40	68.21	78.04	68.50	76.31	68.78	78.89	67.77
MaxLogit	86.68	65.12	87.62	58.30	88.65	50.91	76.74	79.72	84.92	63.59
Energy	83.73	75.04	86.00	68.11	87.87	56.52	72.66	87.66	82.57	71.83
MCM	88.80	57.44	90.09	48.52	88.15	50.40	85.83	57.04	88.20	53.10
ReAct	90.56	52.43	91.87	46.56	92.67	38.79	82.21	75.96	89.33	53.44
DPM-F	94.86	25.08	93.17	31.55	89.89	43.73	87.84	48.90	91.44	37.31
Requires training (or w/ fine-tuning)										
CLIPN-A	94.67	28.75	92.85	31.87	86.93	50.17	87.68	49.49	90.53	40.07
DPM-T	97.66	12.79	94.94	26.89	92.61	31.42	87.65	45.91	93.22	29.25

4.2 Comparison with SOTA

OOD detection on small-scale datasets. In Table 1, we present the results on CIFAR-100. Among the training-free methods, DPM-F achieves the best average performance, with an AUROC of 85.32% and an FPR95 of 57.35%. Further improvements are observed when incorporating lightweight training, with DPM-T increasing AUROC by 6.3% and reducing FPR95 by 18.48% compared to DPM-F. Using the ViT-B-32 backbone, similar trends are shown, DPM-F obtains an AUROC of 84.14%, while training boosts performance markedly to 94.79% AUROC. Considering the performance across various datasets, our method exhibits strong generalization capabilities in OOD detection.

OOD detection on large-scale datasets. ImageNet is a large-scale dataset and the OOD detection performance degrades as explained in [18]. We give a comprehensive OOD evaluation on ImageNet-1K in Table 2. Interestingly, the feature-based method ReAct [35] demonstrates competitive performance compared with the SOTA method CLIPN-A [43]. As a training-free method, ReAct underlines the significance of ID trainset visual information especially on the large-scale dataset. Notably, we only use 10% of the trainset in ImageNet, about 100k images containing 1k categories for ReAct and DPM-F. On average, our training-free method DPM-F with ViT-B-16 achieves 92.78%(+0.73%) AUROC and 29.59%(-13.94%) FPR95. Furthermore, our training-required method

DPM-T adapted on ImageNet with a small sample size can achieve higher AUROC(+2.94%) and FPR95(-8.50%) compared to DPM-F. Similarly, our method using ViT-B-32 achieves 91.44% AUROC (+2.11%) and 29.59% FPR95 (-15.79%) and achieve 1.78% AUROC, 8.06% FPR95 improvement compared to DPM-F.

Table 3: ID val-set: ImageNet val set, **Table 4:** ID val-set: ImageNet covariate shift set, testing on iNaturalist, etc.

OOD	Method	AUROC $\uparrow$	FPR95 $\downarrow$
IN-R	MSP	66.74	80.38
	MCM	73.32	77.79
	DPM-F	78.34	65.08
IN-A	MSP	77.71	84.02
	MCM	80.74	73.57
	DPM-F	86.76	53.73
IN-v2	MSP	66.74	80.38
	MCM	73.32	77.78
	DPM-F	84.84	56.8

ID	Method	AUROC $\uparrow$	FPR95 $\downarrow$
IN-R	MSP	62.65	92.33
	MCM	70.18	89.07
	DPM-F	74.23	76.86
IN-A	MSP	53.35	92.32
	MCM	64.55	88.92
	DPM-F	65.91	88.62
IN-v2	MSP	74.54	76.66
	MCM	83.98	68.09
	DPM-F	87.94	51.88

Robustness to challenging benchmarks. The recent paper [47] takes into account the covariate shift benchmark which focuses on detecting the test data still belonging to an ID category but non-i.i.d. Following [47], we utilize several versions of the ImageNet test set similar to Imagenet-1K but with diverse covariate shifts, including ImageNet-A [7], ImageNet-R [14], ImageNetV2 [33]. These datasets are used as the ID datasets, while OOD samples are sourced from iNaturalist, SUN, Texture, and Places365. The average results are shown in Table 4. Additionally, we employ these datasets as the OOD datasets, while still using the ImageNet-1K as the ID testsets, and the average results on these OOD datasets are shown in Tables 3 and 9. Note that NINCO [2] is an ID-contamination-free OOD test dataset. In both experimental setups, we compare our method with the training-free OOD detection methods MSP and MCM. The results reveal that methods solely based on the original CLIP features, such as MCM, struggled to effectively handle covariate shifts, highlighting the challenges involved in OOD detection with these datasets. In contrast, our method DPM-F showcases significant improvements compared to MCM when dealing with these datasets.

4.3 Ablation Study

Different losses. In Table 5, we present the performance of OOD detection on CIFAR-100 using different combinations of $\mathcal{L}_{org}$, $\mathcal{L}_{VM}$ and $\mathcal{L}_{TM}$. $\mathcal{L}_{org}$ calculate the loss based on the similarity between the original projected global feature $L_v(F^V)$ and $L_t(F^T)$. The combination of $\mathcal{L}_{VM}$ and $\mathcal{L}_{TM}$ achieves the best performance with an AUROC of 91.62% and a FPR95 of 38.87%.

Different training modules. In Table 6, we present the performance of OOD detection on CIFAR-100 using different training modules. The projection lay-

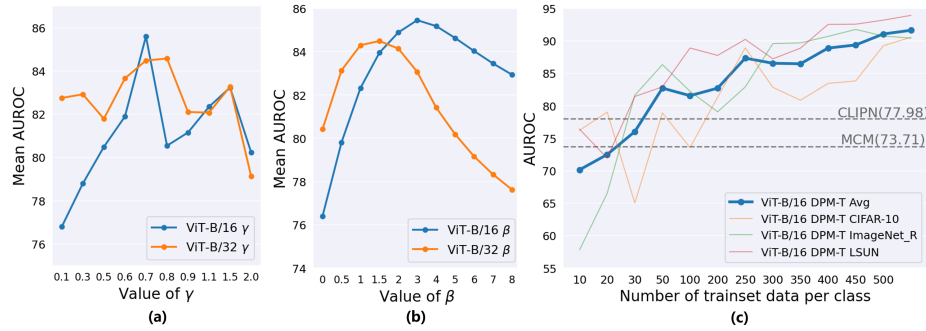


Fig. 3: (a) and (b) Effect of γ and β in DPM-F on CIFAR-100, respectively. (c) Ablation study on CIFAR-100 of DPM-T with ViT-B-16. The gray lines are the average AUROC of CLIPN and MCM.

Table 5: Ablation study of the Loss on CIFAR with ViT-B-16. **Table 6:** Ablation study of training module on CIFAR with ViT-B-32.

$\mathcal{L}_{org}$	$\mathcal{L}_{VM}$	$\mathcal{L}_{TM}$	AUROC $\uparrow$	FPR95 $\downarrow$	Prompt	Projection	DSFA	AUROC $\uparrow$	FPR95 $\downarrow$
✓	×	×	86.65	55.90	×	×	×	84.14	54.26
×	✓	×	86.08	57.15	✓	×	×	84.99	55.69
×	×	✓	61.14	84.87	×	×	✓	76.93	71.27
✓	✓	×	87.23	51.63	✓	✓	×	92.63	38.08
✓	×	✓	84.07	60.52	✓	×	✓	85.12	56.34
×	✓	✓	91.62	38.87	×	✓	✓	90.54	38.95
✓	✓	✓	91.07	39.93	✓	✓	✓	94.79	26.41

ers are one of the critical modules to make DPM-T effective. Furthermore, the combination of all three training modules yields the best overall performance.

Different hyper-parameter for DPM-F. We test the influence of different feature aggregation factors γ for the OOD detection performance on CIFAR-100, as shown in Fig. 3(a). An appropriate value of γ can lead to higher OOD gains. For example, using the ViT-B-16 model and setting γ to 0.7 results in an approximately 8% AUROC improvement compared to that of $\gamma = 0$. We also experiment with the impact of visual modality affinity factor β on the performance of the Dual-Pattern Matching, as shown in Fig. 3(b). Note that the γ in DPM-T is learnable and we only choose the β for different ID datasets.

The effectiveness of Dual-Pattern Matching. We conduct separate experiments to test the performance of using only visual patterns and textual patterns for OOD detection on CIFAR-100, as shown in the upper part in Table 7. We perform visual matching using KL divergence, Euclidean distance, and cosine distance. We also consider using the trainset visual feature F^V prototype as the visual pattern and perform matching using cosine distance (CosFea in the table). Then, we also experiment with combining the score with TM. All experimental results are presented in Table 7, the aggregation of visual matching scores

Table 7: Ablation study of visual matching on CIFAR-100. ED: euclidean distance.

Methods	CIFAR-10		ImageNet-R		LSUN	
	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$
KL	63.91	86.32	61.11	85.69	88.98	42.74
ED	44.24	97.88	75.29	75.66	43.15	98.55
Cos	59.49	92.76	86.69	54.19	55.23	96.48
CosFea	60.40	88.22	92.93	36.87	49.15	97.23
TM	77.23	83.78	76.31	62.99	87.71	67.44
KL + TM	78.72	74.53	78.30	64.44	95.41	23.81
ED + TM	61.54	91.44	84.18	55.33	69.24	90.96
Cos + TM	76.04	79.01	88.49	39.31	84.81	73.05
CosFea + TM	69.36	84.56	91.68	37.16	68.53	93.59

Table 8: Mean AUROC of training-free methods on CIFAR-100 with different numbers of trainset image per category. **Table 9:** ID val-set: ImageNet val set, testing on OOD datasets.

Num	ViT-B-16		ViT-B-32		OOD	Method	AUROC $\uparrow$	FPR95 $\downarrow$
	DPM-F	VM	DPM-F	VM				
5	78.44	63.09	83.64	63.92	NINCO	MSP	72.77	82.99
10	76.84	58.38	83.86	67.42		MCM	73.54	82.39
15	78.99	61.48	82.40	61.56		DPM-F	75.57	76.96
20	83.73	74.59	81.42	59.52		DPM-T	81.37	66.51
50	82.03	70.51	85.68	71.56	IN-v2	MSP	66.74	80.38
500	85.33	76.76	84.14	72.33		MCM	73.32	77.78
						DPM-F	84.84	56.8
						DPM-T	84.76	59.24

using KL divergence and texture matching scores achieves the best OOD performance. The results indicate that using a single modality of matching performs lower than our DPM method, which leverages both visual and textual patterns.

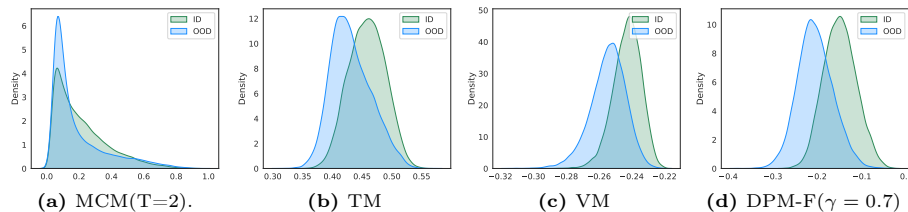
The impact of trainset number of images. We also evaluate the impact of different training set sample sizes for DPM-F and DPM-T model, as shown in Table 8 and Fig. 3(c) separately. Overall, as the number of training set samples increases, the performance shows an upward trend. Furthermore, achieving good OOD performance requires only a small portion of the samples, especially for DPM-F. Using as few as 20 samples per category can achieve performance comparable to using 500 samples. Our DPM-T method achieves performance surpassing CLIPN by using only around 100 samples per category.

The effectiveness of fine-tuning methods. Regarding the fine-tuning approaches for the CLIP model, we compare our DPM-T method with CoOp [56] and CALIP [12] and LoCoOp [29], as shown in Table 10. Due to the specific optimization of our method for OOD detection, we achieved significant performance improvements compared to these fine-tuning methods.

Visualization on the effectiveness of DPM. We showcase OOD detection performance of DPM and MCM on CIFAR-100 (ID) and ImageNet-R (OOD). Since ImageNet-R shares labels across domains, posing challenges for CLIP-

Table 10: Ablation study on different training methods on CIFAR-100 with ViT-B-32.

Method	CIFAR-10		ImageNet-R		LSUN	
	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$
CoOp+MaxLogit	83.87	64.95	91.06	40.53	87.94	58.88
CALIP+MaxLogit	75.18	88.08	67.20	77.75	90.19	65.31
LoCoOp+GMCM	83.26	70.29	57.73	97.02	94.83	29.26
DPM-T	94.11	30.80	93.87	28.54	96.39	19.89

**Fig. 4:** Similarity density of MCM and DPM with ViT-B-16 on CIFAR-100

based models (Fig. 4(a)), DPM enhances performance by incorporating domain-specific features in visual and textual matching (Fig. 4(b)(c)(d)).

5 Conclusion

This paper introduces Dual-Pattern Matching, an efficient adaptation of CLIP for OOD detection. DPM uses a domain-specific feature aggregation module to refine the CLIP image features. Then, DPM stores class-specific text features and vision-text matching prototypes as dual patterns for comparison with test samples. We introduce training-free DPM-F and training-required DPM-T extensions incorporating lightweight adaptation and prompt tuning. DPM outperforms prior CLIP-based approaches in benchmarks by effectively leveraging multimodal information. The proposed DPM provides a simple yet powerful framework to exploit vision-language representations for robust OOD detection.

Limitation. Our method has several limitations. First, DPM accesses a certain number of training samples. Exploring more efficient fine-tuning strategies for DPM, especially in few-shot scenarios, is an area for future improvement. Second, we lack consideration of textual features (*e.g.*, prompt enhancement), which can provide additional context and boost the performance of VLMs.

Acknowledgements

This work was supported by the Natural Science Fund of Hubei Province (Grant 2022CFB823), HUST Independent Innovation Research Fund (Grant 2021XXJS096), and grants from the National Key Lab of MSIPT (Grant 6142113220309) as well as the MoE Key Lab of Image Processing & Intelligent Control, among others.

References

1. Bitterwolf, J., Meinke, A., Hein, M.: Certifiably adversarially robust detection of out-of-distribution data. *Advances in Neural Information Processing Systems* **33**, 16085–16095 (2020)
2. Bitterwolf, J., Müller, M., Hein, M.: In or out? fixing imagenet out-of-distribution detection evaluation. *arXiv preprint arXiv:2306.00826* (2023)
3. Chen, J., Li, Y., Wu, X., Liang, Y., Jha, S.: Atom: Robustifying out-of-distribution detection using outlier mining. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. pp. 430–445. Springer (2021)
4. Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J.: Reproducible scaling laws for contrastive language-image learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 2818–2829 (2023)
5. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3606–3613 (2014)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 248–255. Ieee (2009)
7. Djolonga, J., Yung, J., Tschannen, M., Romijnders, R., Beyer, L., Kolesnikov, A., Puigcerver, J., Minderer, M., D’Amour, A., Moldovan, D., et al.: On robustness and transferability of convolutional neural networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 16458–16468 (2021)
8. Dong, X., Guo, J., Li, A., Ting, W.T., Liu, C., Kung, H.: Neural mean discrepancy for efficient out-of-distribution detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 19217–19227 (2022)
9. Du, X., Wang, Z., Cai, M., Li, Y.: Vos: Learning what you don’t know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197* (2022)
10. Esmaeilpour, S., Liu, B., Robertson, E., Shu, L.: Zero-shot out-of-distribution detection based on the pre-trained model clip. In: *Proceedings of the AAAI conference on artificial intelligence*. vol. 36, pp. 6568–6576 (2022)
11. 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* pp. 1–15 (2023)
12. Guo, Z., Zhang, R., Qiu, L., Ma, X., Miao, X., He, X., Cui, B.: Calip: Zero-shot enhancement of clip with parameter-free attention. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 37, pp. 746–754 (2023)
13. Hendrycks, D., Basart, S., Mazeika, M., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132* (2019)
14. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical analysis of out-of-distribution generalization. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 8340–8349 (2021)
15. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations* (2017)
16. Hsu, Y.C., Shen, Y., Jin, H., Kira, Z.: Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In: *Proceedings*

- of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10951–10960 (2020)
17. Huang, R., Geng, A., Li, Y.: On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems* **34**, 677–689 (2021)
 18. Huang, R., Li, Y.: Mos: Towards scaling out-of-distribution detection for large semantic space. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8710–8719 (2021)
 19. Joseph, K.J., Khan, S., Khan, F.S., Balasubramanian, V.N.: Towards open world object detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 5830–5840 (June 2021)
 20. Koner, R., Sinhamahapatra, P., Roscher, K., Günnemann, S., Tresp, V.: Ood-former: Out-of-distribution detection transformer. *arXiv preprint arXiv:2107.08976* (2021)
 21. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
 22. Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., Yan, J.: Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208* (2021)
 23. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: *International Conference on Learning Representations* (2018)
 24. Liang, V.W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.Y.: Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems* **35**, 17612–17625 (2022)
 25. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems* **33**, 21464–21475 (2020)
 26. Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., Li, Y.: Delving into out-of-distribution detection with vision-language representations. *Advances in Neural Information Processing Systems* **35**, 35087–35102 (2022)
 27. Ming, Y., Li, Y.: How does fine-tuning impact out-of-distribution detection for vision-language models? *arXiv preprint arXiv:2306.06048* (2023)
 28. Ming, Y., Sun, Y., Dia, O., Li, Y.: How to exploit hyperspherical embeddings for out-of-distribution detection? *arXiv preprint arXiv:2203.04450* (2022)
 29. Miyai, A., Yu, Q., Irie, G., Aizawa, K.: Locoop: Few-shot out-of-distribution detection via prompt learning. In: *Thirty-Seventh Conference on Neural Information Processing Systems* (2023)
 30. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 427–436 (2015)
 31. Papadopoulos, A.A., Rajati, M.R., Shaikh, N., Wang, J.: Outlier exposure with confidence control for out-of-distribution detection. *Neurocomputing* **441**, 138–150 (2021)
 32. 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)
 33. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: *International conference on machine learning*. pp. 5389–5400. PMLR (2019)

34. Sun, X., Hu, P., Saenko, K.: Dualcoop: Fast adaptation to multi-label recognition with limited annotations. In: *Advances in Neural Information Processing Systems* (2022)
35. Sun, Y., Guo, C., Li, Y.: React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems* **34**, 144–157 (2021)
36. Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-distribution detection with deep nearest neighbors. *arXiv preprint arXiv:2204.06507* (2022)
37. Tao, L., Du, X., Zhu, X., Li, Y.: Non-parametric outlier synthesis. *arXiv preprint arXiv:2303.02966* (2023)
38. Thulasidasan, S., Chennupati, G., Bilmes, J.A., Bhattacharya, T., Michalak, S.: On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems* **32** (2019)
39. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 8769–8778 (2018)
40. Vojř, T., Šochman, J., Aljundi, R., Matas, J.: Calibrated out-of-distribution detection with a generic representation. In: *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. pp. 4509–4518. IEEE (2023)
41. Wan, Q., Wang, S., Xiang, X.: A simple unknown-instance-aware framework for open-set object detection. In: *2023 13th International Conference on Information Science and Technology (ICIST)*. pp. 586–593. IEEE (2023)
42. Wang, H., Li, Z., Feng, L., Zhang, W.: Vim: Out-of-distribution with virtual-logit matching. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4921–4930 (2022)
43. Wang, H., Li, Y., Yao, H., Li, X.: Clipn for zero-shot ood detection: Teaching clip to say no. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 1802–1812 (2023)
44. Xiang, X., Tan, Y., Wan, Q., Ma, J., Yuille, A., Hager, G.D.: Coarse-to-fine incremental few-shot learning. In: *European Conference on Computer Vision*. pp. 205–222. Springer (2022)
45. Xiang, X., Zhang, Z., Chen, X.: Curricular-balanced long-tailed learning. *Neurocomputing* **571**, 127121 (2024)
46. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: *2010 IEEE computer society conference on computer vision and pattern recognition*. pp. 3485–3492. IEEE (2010)
47. Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., Peng, W., Wang, H., Chen, G., Li, B., Sun, Y., et al.: Openood: Benchmarking generalized out-of-distribution detection. *arXiv preprint arXiv:2210.07242* (2022)
48. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783* (2021)
49. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365* (2015)
50. Yu, Q., Aizawa, K.: Unsupervised out-of-distribution detection by maximum classifier discrepancy. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 9518–9526 (2019)
51. Zhang, J., Fu, Q., Chen, X., Du, L., Li, Z., Wang, G., Han, S., Zhang, D., et al.: Out-of-distribution detection based on in-distribution data patterns memorization with

- modern hopfield energy. In: The Eleventh International Conference on Learning Representations (2022)
52. 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: European Conference on Computer Vision. pp. 493–510. Springer (2022)
 53. Zhang, Z., Xiang, X.: Decoupling maxlogit for out-of-distribution detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3388–3397 (June 2023)
 54. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence* **40**(6), 1452–1464 (2017)
 55. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII. pp. 696–712. Springer (2022)
 56. 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)
 57. Zhou, Q., Zhang, Z., Xiang, X., Wang, K., Wu, Y., Li, Y.: Enhancing the general agent capabilities of low-parameter llms through tuning and multi-branch reasoning. *arXiv preprint arXiv:2403.19962* (2024)