Enriching Category Representations with LLMs Towards Robust Zero-Shot OOD Detection

Dian Chao[1], Yuxuan Zhang[1], Luping Zhou[2], and Yang Yang[1] (✉)

[1] Nanjing University of Science and Technology, Nanjing 210000, China
{chaodian,xuan_yuzhang,yyang}@njust.edu.cn
[2] University of Sydney, Sydney 2006, Australia luping.zhou@sydney.edu.cn

Abstract. Recent advancements in foundation models, particularly Visual-Language Models (VLMs) have enabled effective zero-shot Out-of-distribution (OOD) detection. Existing methods attempt to generate the names of OOD classes similar to in-distribution (ID) classes to explore the textual space of VLMs. However, they fail to integrate relevant ID information to reveal specific OOD features, thus limiting the distinction between ID and OOD classes. To address this issue, we propose a simple yet effective zero-shot OOD detection approach incorporating a specific semantic text generation strategy and a new regionally enhanced semantic OOD scoring function. In detail, we employ meticulously designed prompts to generate challenging OOD label texts using Large Language Models (LLMs). Subsequently, the specific semantic text generation strategy leverages LLMs to capture fine-grained textual representations of both ID and OOD classes. Additionally, the regionally enhanced semantic OOD score is formulated by adjusting the confidence of ID classes to improve OOD detection. Experiments demonstrate that our method achieves state-of-the-art (SOTA) performance on multiple OOD detection benchmarks. The code is available at repository.

Keywords: Out-of-distribution Detection · Zero-shot Learning · Visual-Language Models.

1 Introduction

With the continuous development of deep learning and foundation models, the research community has shifted from traditional i.i.d. assumptions towards open-world scenarios. Consequently, traditional models exhibit performance degradation on OOD data [1,2,3,4,5,29,30,52]. In response, OOD detection has become essential for identifying and rejecting invalid inputs and ensuring safety. This capability is particularly crucial in high-stakes domains such as autonomous driving [6,7] and medical diagnostics [8].

To address these challenges, existing methods can be categorized into two historical stages: 1) vision-only methods [9,10,11,12,13,31]. and 2) vision-language methods. The vision-only approaches primarily focus on utilizing external OOD

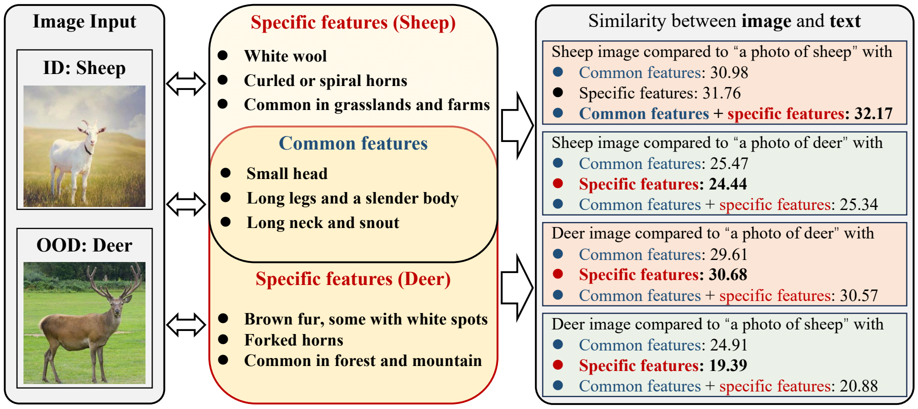


Image Input
ID: Sheep
OOD: Deer
Specific features (Sheep)
White wool
Curled or spiral horns
Common in grasslands and farms
Common features
Small head
Long legs and a slender body
Long neck and snout
Specific features (Deer)
Brown fur, some with white spots
Forked horns
Common in forest and mountain
Similarity between image and text
Sheep image compared to "a photo of sheep" with
Common features: 30.98
Specific features: 31.76
Common features + specific features: 32.17
Sheep image compared to "a photo of deer" with
Common features: 25.47
Specific features: 24.44
Common features + specific features: 25.34
Deer image compared to "a photo of deer" with
Common features: 29.61
Specific features: 30.68
Common features + specific features: 30.57
Deer image compared to "a photo of sheep" with
Common features: 24.91
Specific features: 19.39
Common features + specific features: 20.88

Fig. 1. Image-Text Similarity of Sheep and Deer: Comparing Common and Specific Features. Specific features refer to the unique characteristics of sheep or deer, while common features represent the shared traits. The right side shows the similarity scores between images of sheep/deer and text descriptions that include 'common', 'specific', or 'common+specific' features. For each image, we aim to maximize similarity with the corresponding class description while minimizing similarity with descriptions of unrelated classes.

images to enhance model robustness or exploring uncertainty in visual representations across varying distributions, without taking into account the potential benefits introduced by textual information [14,15]. With the development of foundation models, VLMs [16] exhibit strong generalization capabilities after being trained on large-scale image-text pairs. In recent years, an increasing number of works [17,18] focus on leveraging textual modality features for OOD detection using VLMs. These approaches demonstrate superior performance compared to previous OOD detection methods. However, these methods primarily utilize ID class names and lack comprehensive exploitation of the textual modality. Recent works have begun to explore more extensive information from the textual modality. Several approaches [14,15] endeavor to generate OOD class names using resources such as WordNet [19], while others [20] leverage LLMs [21] to generate semantic descriptions of ID classes for zero-shot OOD detection.

However, existing methods tend to overlook the integration of ID information necessary for capturing distinctive OOD textual features. We argue that relying solely on textual features derived from OOD names or ID class descriptors is insufficient for effectively distinguishing hard OOD instances. Leveraging VLMs' ability to align textual and visual features, we can guide the model to focus on regions unique to hard OOD instances. To verify the intention, we first analyze the influence of common and specific features, as depicted in Fig.1. Possibly, here give an example of a common feature and an example of a specific feature by combining descriptive terms (e.g., "a photo of a sheep with white wool, commonly found in grasslands and farms") and computing similarity with both sheep and deer images, we observe that common features lead to misclassification.

In contrast, more specific features help reduce it. Acquiring such fine-grained textual descriptions of OOD classes can significantly enhance OOD detection performance. Unfortunately, large lexical databases like WordNet, while useful for constructing categorical relationships, lack the contextual specificity needed to capture specific features for each category. With the advancement of LLMs trained on extensive text, these models have acquired broad knowledge and the capability to analyze relationships and distinctions between categories. This work we harness the power of LLMs to generate fine-grained textual descriptions.

Therefore, we propose a simple yet effective zero-shot OOD detection approach that utilizes LLMs to generate names for hard OOD classes resembling ID classes while systematically excluding synonyms and near-synonyms through similarity calculations. To enhance VLMs' OOD detection capacities, fine-grained descriptions are generated by simultaneously considering both ID and OOD class names. Specifically, this work uses LLMs to generate descriptions for ID classes. Subsequently, LLMs are also employed to generate OOD classes that are prone to be misclassified as the given ID classes, along with specific features that distinguish these hard OOD classes from their ID counterparts. This approach aims to maximize the separation of textual feature spaces between hard OOD and ID classes. Furthermore, the impact of shared features between OOD and ID classes on OOD detection performance is analyzed using information entropy, demonstrating that common features adversely affect OOD detection. To address this, we propose a novel scoring method that adjusts the confidence of ID samples based on their similarity to hard OOD classes. Extensive experiments demonstrate that our method achieves SOTA performance across multiple datasets. This approach enhances the model's performance in hard OOD detection tasks and exhibits strong generalization capabilities. In summary, our key contributions are as follows:

- We further explore the textual space at the regional feature level using semantic texts generated by LLMs, aiming to identify discriminative regional features between ID and OOD counterparts and maximize their separation.
- We analyze the influence of common and specific features on OOD detection. Moreover, we propose a regionally enhanced semantic OOD score, adjusting ID confidence based on similarity to synthesized OOD classes.
- The effectiveness of the proposed method is validated across diverse settings, encompassing both simple and challenging OOD tasks. Experimental results indicate that this approach achieves SOTA performance across multiple OOD detection benchmarks.

2 Related Work

Traditional OOD Detection is typically categorized into two types: training-time regularization [22,23,24,25,26,28] and post hoc methods [2,12,13,27,32,33,51]. Training-time regularization methods assume that a subset of OOD data is accessible during model training. CSI [22] enhances the OOD detector through the application of contrastive learning. MOS [23] pre-groups all categories and introduces an additional class to each group, redesigning the loss function for training. VOS [24] improves energy scores by generating virtual anomalies. LogitNorm [25] offers an alternative to cross-entropy loss by separating the influence of the logit norm from the training process. CIDER [26] improves OOD detection performance by optimizing contrastive loss.

Post hoc methods do not alter the model's parameters; instead, they typically focus on designing an OOD score. MSP [2] utilizes the highest predicted softmax probability as the OOD score. ODIN [27] refines MSP by applying input perturbations and rescaling the logits. Energy [13] introduces the use of an energy function [34] to quantify OOD. Mahalanobis [12] calculates the OOD score based on the minimum Mahalanobis distance between the feature and the centroids of each class. GradNorm [32] develops the OOD score by utilizing the gradient space. ViM [35] integrates the norm of feature residuals with the principal space created by training features and the original logits to determine the degree of OOD-ness. KNN [33] explores the effectiveness of non-parametric nearest-neighbor distances for identifying OOD samples.

OOD Detection based on VLMs has been developed using CLIP [16] as the foundation, leveraging its powerful vision-language alignment capabilities. MCM [17] introduced this approach by utilizing maximum softmax probabilities to assess the similarity of images to known classes, thereby identifying OOD images. ZOC [36] train image decoders for extracting textual information from images. CLIPN [18] proposes constructing negative sample pairs and conducting pre-training to learn a 'no' concept for each class. MMOOD [20] propose using LLMs to generate additional descriptive terms for ID classes to enrich textual semantic information. Recent studies [14,15] have explored methods to leverage VLMs' zero-shot inference capability by generating OOD categories through various approaches, aiming to represent potential OOD scenarios. Specifically, EOE [14] utilizes LLMs to generate potential outlier class and designs an outlier penalty function to detect OOD samples. NegLabel [15] acquisition utilizes WordNet to gather a diverse set of OOD category names, complemented by a scoring function to identify the OOD class with low similarity to current IDs.

Large Language Models such as GPT-3 [37], LLaMA-3 [38], GPT-4 [39], are leading advancements in natural language processing. These models are trained on massive datasets with parameters ranging from hundreds of billions to trillions. LLMs represent significant advancements in natural language processing, pushing boundaries in language understanding, generation, and adaptation across various domains. Given LLMs' broad knowledge base, they are instrumental in providing similarities and differences among categories akin to ID.

3 METHODOLOGY

This section details the proposed approach. Section 3.1 defines the notation and outlines the problem. Section 3.2 introduces a method for generating outliers and fine-grained features by leveraging LLMs to augment class descriptions. The complete framework is depicted in Fig.3. In Section 3.3, a novel OOD detection scoring function is presented, which clusters ID and OOD categories.

3.1 Notation and Preliminary

Without loss of generality, assume that we have n images which are denoted as $\boldsymbol{X} = \{x_1, \cdots, x_n\}$. The ID class names set $\boldsymbol{Y}^{(\mathrm{id})} = \{y_1^{(\mathrm{id})}, \cdots, y_c^{(\mathrm{id})}\}$ is also available, where c denotes the number of class names. The goal of OOD detection is to determine whether an image $x \in \boldsymbol{X}$ belongs to the ID class $\boldsymbol{Y}^{(\mathrm{id})}$ or not.

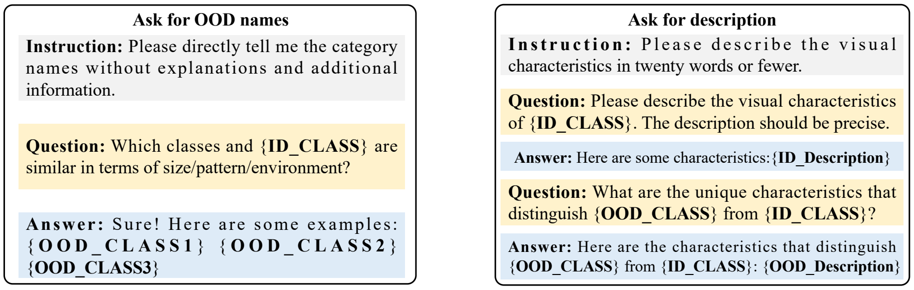
We prompt LLMs to generate OOD class names set to assist the OOD detection task. Additionally, extra information for both ID and OOD classes is generated to better align text and images. The descriptions for ID classes are denoted by $\boldsymbol{D}^{(\mathrm{id})}$, and the OOD class names set and their descriptions are denoted by $\boldsymbol{Y}^{(\mathrm{ood})}$ and $\boldsymbol{D}^{(\mathrm{ood})}$, respectively. Furthermore, a pre-trained model is used to encode text including class name and description and image as feature, and then decide whether an image belongs to the ID class names set. Specifically, we use $\phi(\cdot)$ and $\psi(\cdot)$ to denote the image and text encoder, respectively. For an image $\boldsymbol{x}$ and a text $\boldsymbol{t}$, their features can be calculated by:

$$\boldsymbol{u} = \phi(\boldsymbol{x}), \ \boldsymbol{v} = \psi(\boldsymbol{t}), \tag{1}$$

where $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{R}^d$ denote the image and text features, d denotes the feature dimension, respectively. The text input can be a class name or description.

Based on the features of the given image and text information, we can design an evaluation function to decide whether the image belongs to the ID class or not. For a comprehensive list of notations, refer to the appendix A.1.

3.2 Specific Semantic Text Generation Strategy



Ask for OOD names

Instruction: Please directly tell me the category names without explanations and additional information.

Question: Which classes and {**ID_CLASS**} are similar in terms of size/pattern/environment?

Answer: Sure! Here are some examples: {**OOD_CLASS1**} {**OOD_CLASS2**} {**OOD_CLASS3**}

(*a*) OOD class generation.

Ask for description

Instruction: Please describe the visual characteristics in twenty words or fewer.

Question: Please describe the visual characteristics of {**ID_CLASS**}. The description should be precise.

Answer: Here are some characteristics: {**ID_Description**}

Question: What are the unique characteristics that distinguish {**OOD_CLASS**} from {**ID_CLASS**}?

Answer: Here are the characteristics that distinguish {**OOD_CLASS**} from {**ID_CLASS**}: {**OOD_Description**}

(*b*) Description generation.

Fig. 2. (a) The prompt queries to obtain OOD class names similar to ID classes, including instruction, question, and model response examples. (b) The prompt queries to obtain descriptive words for ID and OOD classes, including instruction, question, and model response examples.

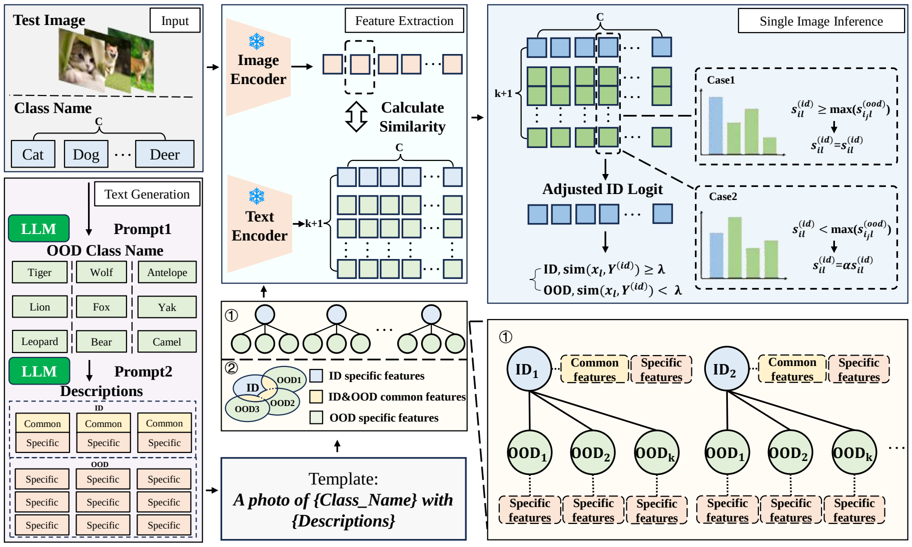


Fig. 3. The main framework of our model. LLMs are employed to generate OOD class names and descriptions for ID samples, following a fixed template. Feature vectors are then extracted using frozen encoders. Finally, the OOD detection process for a single image is demonstrated. In the enlarged section labeled ①, a detailed depiction is provided of how the relationship between an ID class and its corresponding hard OOD is constructed, along with their associated textual descriptions.

LLMs are adopted to generate additional textual information to support the OOD detection task. Beyond generating OOD class names, LLMs are also used to create nuanced textual descriptions for both ID and OOD class names.

LLMs are first utilized to generate OOD class names. To enhance the discriminative ability of the model, the generated OOD classes are required to closely resemble the ID classes. For example, if the ID class name is "cat", an OOD class name like "tiger" is preferred over unrelated options such as "book". To achieve this goal, prompts are refined to drive LLMs to generate target OOD class names by exploring the pivotal properties including size, pattern, and environment. The interaction process with LLMs is shown in Fig.2 (a).

Formally, k OOD class names are generated for each ID class name, i.e., $\forall \boldsymbol{y}_i^{(\mathrm{id})} \in \boldsymbol{Y}^{(\mathrm{id})}$, we generate OOD class names set $\{\boldsymbol{y}_{i1}^{(\mathrm{ood})}, \cdots, \boldsymbol{y}_{iK_i}^{(\mathrm{ood})}\}$. $\boldsymbol{Y}^{(\mathrm{ood})}$ are utilized to denote the whole OOD class names set, which is defined as follows:

$$\boldsymbol{Y}^{(\mathrm{ood})} = \bigcup_{i=1}^{c} \{\boldsymbol{y}_{i1}^{(\mathrm{ood})}, \cdots, \boldsymbol{y}_{iK_i}^{(\mathrm{ood})}\}, \tag{2}$$

where K_i denotes the number of generated OOD class names and its value is dependent on the output of LLMs.

Since the strong knowledge capacity of the LLMs, it is necessary to filter out some OOD classes [15] that are too similar to ensure the distinguishability of the subsequent description generation. For a detailed explanation, please refer to appendix B.1. To achieve this, features for both ID classes and their corresponding OOD class names are first extracted. Specifically, a pre-trained CLIP model is utilized to extract the textual features for a given ID class and its associated OOD class names:

$$\boldsymbol{v}_i^{(\mathrm{id})} = \psi(\boldsymbol{y}_i^{(\mathrm{id})}), \tag{3}$$

$$\forall j \in \{1, \cdots, k\}, \boldsymbol{v}_{ij}^{(\mathrm{ood})} = \psi(\boldsymbol{y}_{ij}^{(\mathrm{ood})}). \tag{4}$$

Then, the similarity between ID class and OOD class names is calculated

$$\forall j \in \{1, \cdots, k\}, s_{ij} = \frac{[\boldsymbol{v}_i^{(\mathrm{id})}]^\top \boldsymbol{v}_{ij}^{(\mathrm{ood})}}{\|\boldsymbol{v}_i^{(\mathrm{id})}\| \|\boldsymbol{v}_{ij}^{(\mathrm{ood})}\|}. \tag{5}$$

According to similarity, for each ID class, the top-k OOD class names with the lowest similarity scores are selected to construct pairs, where $k \leq \min\{K_i\}_{i=1}^c$. This process results in the filtered OOD class name set:

$$\hat{\boldsymbol{Y}}^{(\mathrm{ood})} = \bigcup_{i=1}^{c} \{\boldsymbol{y}_{il_1}^{(\mathrm{ood})}, \cdots, \boldsymbol{y}_{il_k}^{(\mathrm{ood})}\}, \tag{6}$$

where $|\hat{\boldsymbol{Y}}^{(\mathrm{ood})}| = ck$.

After obtaining the ID class name $\boldsymbol{Y}^{(\mathrm{id})}$ and the filtered OOD class names $\hat{\boldsymbol{Y}}^{(\mathrm{ood})}$, descriptions for each class are generated. Given the high similarity between an ID class name and its corresponding OOD class names, a novel strategy is devised to generate distinctive descriptions for each ID and OOD class.

For each ID class name and its paired similar OOD class names, we prompt LLMs to generate the description $\boldsymbol{D}^{(\mathrm{id})} = \{\boldsymbol{d}_1^{(\mathrm{id})}, \cdots, \boldsymbol{d}_c^{(\mathrm{id})}\}$ that characterizes the ID class name concisely and precisely. For example, for the ID class name "sheep", the generated descriptions might include "white wool", "long neck and snout". These descriptions may overlap with features of similar OOD classes, for example, the OOD class "deer" similar to ID class "sheep", could also be described as "long neck and snout". Next, we prompt LLMs to generate concise and precise descriptions for paired OOD class names. In appendix B.2, we analyze the impact of common and specific features on OOD detection, demonstrating that only specific features can enhance OOD detection. To ensure these descriptions highlight unique properties and avoid overlapping with the ID class, the LLMs are explicitly instructed to focus on distinguishing features in their generated descriptions $\boldsymbol{D}^{(\mathrm{ood})} = \bigcup_{i=1}^{c} \{\boldsymbol{d}_{i1}^{(\mathrm{ood})}, \cdots, \boldsymbol{d}_{ik}^{(\mathrm{ood})}\}$. The prompt for description generation is given in Fig.2 (b).

3.3 Regionally Enhanced Semantic OOD Score

A novel method is proposed to compute the ID similarity score to complete the OOD detection task.

For any image $\boldsymbol{x}_l$ and ID class name $\boldsymbol{y}_i^{(\mathrm{id})}$, we utilize the OOD class names $\{\hat{\boldsymbol{y}}_{i1}^{(\mathrm{ood})}, \cdots, \hat{\boldsymbol{y}}_{ik}^{(\mathrm{ood})}\}$ corresponding to $\boldsymbol{y}_i^{(\mathrm{id})}$, the ID description $\boldsymbol{d}_i^{(\mathrm{id})}$ corresponding to $\boldsymbol{y}_i^{(\mathrm{id})}$, and the OOD descriptions $\{\boldsymbol{d}_{i1}^{(\mathrm{ood})}, \cdots, \boldsymbol{d}_{ik}^{(\mathrm{ood})}\}$ to obtain the confidence score of image and ID class. Since there is a one-to-one correspondence between the class name and their descriptions, we first combine them using the following prompt to obtain an input text:

$$\boldsymbol{t} = \text{“}\textit{A photo of \{CLASS\_NAME\} with \{DESCRIPTION\}.}\text{”}$$

By respectively substituting $CLASS\_NAME$ and $DESCRIPTION$ with the ID class name and its description, we obtain $\boldsymbol{t}_i^{(\mathrm{id})}$. Similar operations are used to obtain $\{\boldsymbol{t}_{i1}^{(\mathrm{ood})}, \cdots, \boldsymbol{t}_{ik}^{(\mathrm{ood})}\}$. Then, according to image $\boldsymbol{x}_l$, text $\boldsymbol{t}_i^{(\mathrm{id})}$ and $\{\boldsymbol{t}_{i1}^{(\mathrm{ood})}, \cdots, \boldsymbol{t}_{ik}^{(\mathrm{ood})}\}$, we can calculate features by using:

$$\boldsymbol{u}_l = \phi(\boldsymbol{x}_l), \tag{7}$$

$$\boldsymbol{v}_i^{(\mathrm{id})} = \psi(\boldsymbol{t}_i^{(\mathrm{id})}), \tag{8}$$

$$\forall j \in \{1, \cdots, k\}, \boldsymbol{v}_{ij}^{(\mathrm{ood})} = \psi(\boldsymbol{t}_{ij}^{(\mathrm{ood})}). \tag{9}$$

Then, we can calculate the similarity by:

$$s_{il}^{(\mathrm{id})} = \frac{\boldsymbol{u}_l^\top \boldsymbol{v}_i^{(\mathrm{id})}}{\|\boldsymbol{u}_l\| \|\boldsymbol{v}_i^{(\mathrm{id})}\|}, \tag{10}$$

$$\forall j \in \{1, \cdots, k\}, s_{i_j l}^{(\mathrm{ood})} = \frac{\boldsymbol{u}_l^\top \boldsymbol{v}_{ij}^{(\mathrm{ood})}}{\|\boldsymbol{u}_l\| \|\boldsymbol{v}_{ij}^{(\mathrm{ood})}\|}. \tag{11}$$

Based on the similarity $s_{il}^{(\mathrm{id})}$ and $\{s_{i_j l}^{(\mathrm{ood})}\}_{j=1}^k$, a similarity strategy is proposed based on the rectification degree α. Intuitively, when the model is more inclined to classify an image as belonging to an OOD class, the confidence in the ID class should decrease. Formally, an amended similarity is defined as:

$$\hat{s}_{il}^{(\mathrm{id})} = \begin{cases} \alpha s_{il}^{(\mathrm{id})} & \text{if } s_{il}^{(\mathrm{id})} \leq \max_{j=1}^{k} \{s_{i_j l}^{(\mathrm{ood})}\}, \\ s_{il}^{(\mathrm{id})} & \text{otherwise}, \end{cases} \tag{12}$$

where $0 < \alpha < 1$.

For now, we obtain the amended similarity score between an image and all ID class names. Similar to MCM [17], normalized confidence is used to determine whether an image belongs to an ID class. Specifically, the confidence is calculated using the following formula:

$$\mathtt{sim}(\boldsymbol{x}_l, \boldsymbol{t}^{(\mathrm{id})}) = \max_{i=1}^{c} \frac{e^{s_{il}^{(\mathrm{id})}/\tau}}{\sum_{j=1}^{c} e^{s_{jl}^{(\mathrm{id})}/\tau}}, \tag{13}$$

where τ is the temperature coefficient. Then, if the confidence score is larger than a threshold parameter λ, $\boldsymbol{x}_l$ is predicted as belonging to ID classes. Otherwise, $\boldsymbol{x}_l$ belongs to OOD classes. The detailed algorithm is provided in appendix A.2.

Table 1. The OOD performance (%) ImageNet-1k as the ID dataset. The best results are highlighted in **bold**, and the second-best results are in underlined.

Methods	iNaturalist		SUN		Places		Textures		Average	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Requires training (w. fine-tuning)										
MSP	87.44	58.36	79.73	73.72	79.67	74.41	79.69	71.93	81.63	69.61
ODIN	94.64	30.22	87.17	54.04	85.54	55.06	87.85	57.61	88.80	47.75
Energy	95.33	26.12	92.66	35.97	91.41	39.87	86.76	57.61	91.54	39.89
GradNorm	72.56	81.50	72.86	82.00	73.70	80.41	70.26	79.36	72.35	80.82
ViM	93.16	32.19	87.19	54.01	83.75	60.67	87.18	53.94	87.82	50.20
KNN	94.52	29.17	92.67	35.62	91.02	39.61	85.67	64.35	90.97	42.19
VOS	94.62	28.99	92.57	36.88	91.23	38.39	86.33	61.02	91.19	41.32
NPOS	96.19	16.58	90.44	43.77	89.44	45.27	88.80	46.12	91.22	37.93
ZOC	86.09	87.30	81.20	81.51	83.39	73.06	76.46	98.90	81.79	85.19
CLIPN	95.27	23.94	93.93	26.17	92.28	33.45	90.93	40.83	93.10	31.10
Zero-shot (w/o. fine-tuning)										
Mahalanobis	55.89	99.33	59.94	99.41	65.96	98.54	64.23	98.46	61.50	98.94
Energy	85.09	81.08	84.24	79.02	83.38	75.08	65.56	93.65	79.57	82.21
MCM	94.59	32.20	92.25	38.80	90.31	46.20	86.12	58.50	90.82	43.93
MMOOD	95.54	22.88	92.60	34.29	89.87	41.63	87.71	52.02	91.43	37.71
EOE	97.52	12.29	95.73	20.40	92.95	30.16	85.64	57.53	92.96	30.09
NegLabel	**99.49**	**1.91**	95.49	20.53	91.64	35.59	90.22	43.56	94.21	25.40
Ours	98.59	6.03	**96.52**	**18.72**	**93.13**	**28.86**	**92.22**	**39.15**	**95.12**	**23.19**

4 EXPERIMENTS

4.1 Datasets and Metrics

Datasets. In this paper, we evaluate the effectiveness of the proposed methods under two different settings. First, we consider ImageNet-1K [40] as the ID dataset and use iNaturalist [41], SUN [42], Places [43], and Texture [44] as the OOD datasets, following the MCM [17]. Simultaneously, consistent with the settings of works, we use a subset of ImageNet-1k and the Waterbirds dataset [45] as ID datasets. Moreover, We further utilize a distinct subset of ImageNet-1K along with the Spurious OOD dataset [46] as out-of-distribution datasets to evaluate hard OOD detection.
Metrics. Following the setting of prior researches [14,15,17,18], we utilize two metrics: (1) the area under the ROC curve (AUROC), and (2) the false positive rate at 95% true positive rate (FPR95) for OOD samples.

4.2 Compared Methods

We compare our approach with the current SOTA OOD detection methods, encompassing both zero-shot and fine-tuned models. Among fine-tuned models, we evaluate MSP [2], ODIN [27], Energy [13], GradNorm [32], ViM [35], KNN [33], VOS [24], NPOS [47], CLIPN [18], and ZOC [36]. For zero-shot models, we consider MCM [17] along with post-hoc methods applied to the CLIP architecture, including Mahalanobis [12] and Energy [13] as additional baselines, and

MMOOD [20], NegLabel [15], and EOE [14], which enhance category representations by incorporating textual descriptions. Notably, CLIPN [18] utilizes the large-scale CC-3M dataset [48] for additional pre-training of the text encoder.

4.3 Implementation Details

We utilize CLIP[16] as the backbone of our framework, incorporating ViT-B/16 as the image encoder and a masked self-attention transformer as the text encoder. Pre-trained weights for CLIP are adopted from OpenAI. Additionally, for LLMs, we employ LLaMA-3-8B[38], using pre-trained weights provided by Meta. In the experiments, unless otherwise specified, we use $k = 3$ to generate OOD classes corresponding to each ID class and set certification degree α to 0.8. We select the threshold value of λ when 95% of the ID samples are correctly classified and $T = 1$ as the temperature, following the standard practice [17,49]. The configuration of the experimental environment is provided in the appendix A.3.

4.4 Performance Comparison

OOD Detection on Large-Scale Datasets. We use ImageNet-1k as the ID dataset and iNaturalist, SUN, Places, and Texture as the OOD datasets. Table 1 compares our approach with the latest SOTA methods, including both training-based and zero-shot inference methods. Our method achieves SOTA performance on the ImageNet-1k benchmark and surpasses a range of methods that employ fine-tuning for OOD detection, demonstrating the robust zero-shot OOD detection capabilities of CLIP. Furthermore, compared with traditional zero-shot OOD methods including Mahalanobis, Energy, and MCM, approaches like MMOOE, EOE, and NegLabel, which further explore textual features, achieve superior performance. Building upon these methods, our approach delves deeper into the specific textual features of OOD classes, resulting in outstanding performance across multiple datasets. Additionally, the OOD classes constructed for each ID, even if they do not include the exact class names of the OOD samples encountered, provide an expanded feature space that facilitates matching OOD samples. It is noteworthy that our method is slightly outperformed by NegLabel on the iNaturalist dataset. This is because NegLabel generates a large number of OOD class names, and the iNaturalist dataset contains a substantial number of plant species, among which these generated OOD class names are included. The performance results obtained using various VLMs backbones are included in the appendix C.1.

OOD Detection on Hard OOD Datasets. To further demonstrate the effectiveness of the proposed approach, we conduct additional evaluations under two hard OOD conditions: 1) semantically hard OOD and 2) spurious OOD, as shown in Table 2. In detail, semantically hard OOD refers to OOD samples that are semantically similar to the ID samples; for this, we use ImageNet-10 and ImageNet-20 as the ID and OOD datasets, respectively, and vice versa. Spurious OOD refers to OOD samples that have false correlations with the ID samples,

Table 2. Zero-shot OOD detection performance on hard OOD detection tasks.

	ID:ImageNet-10 OOD:ImageNet-20		ID:ImageNet-20 OOD:ImageNet-10		ID:Waterbirds OOD:Spurious OOD		Average	
Methods	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Mahalanobis	90.71	51.46	90.41	37.50	**99.55**	**2.21**	93.56	30.39
Energy	97.94	10.30	97.37	16.40	97.16	7.76	97.49	11.49
MCM	98.71	5.00	98.09	12.91	93.30	14.45	96.70	11.12
MMOOD	98.77	4.20	98.26	9.24	98.62	4.56	98.55	6.00
EOE	99.09	4.20	98.10	13.93	97.69	6.18	98.29	8.10
NegLabel	98.86	5.10	98.81	4.60	94.67	9.50	97.45	6.40
Ours	**99.32**	**1.10**	**99.23**	**1.40**	99.09	4.30	**99.21**	**2.27**

such as the spurious correlation between habitats and bird species. The results indicate that even under more challenging conditions, the proposed method consistently enhances OOD detection performance, achieving an average improvement of 3.73% in FPR95 and 0.66% in AUROC compared to the current SOTA methods. Specifically, on the task where ImageNet-10 serves as the ID dataset and ImageNet-20 as OOD, our method improves FPR95 by 3.10% and AUROC by 0.29%. When the roles are reversed, we observe improvements of 3.20% in FPR95 and 0.42% in AUROC. These results highlight the superior performance of our method in semantically hard OOD detection.

4.5 Ablation Studies

Score Functions. To verify our method on various OOD detection score functions, we have considered several zero-shot OOD methods, as shown in Table 3. We denote the scores before applying our method as MCM, Energy, and MaxLogit, and the scores after applying our method as MCM_{our}, $\text{Energy}_{\text{our}}$, and $\text{MaxLogit}_{\text{our}}$. This approach allows us to validate the impact of MCM scores on the experimental results in the ablation study of our main method. Specifically, after adjusting the ID confidence using our method, we perform OOD detection using the MCM, Energy, and MaxLogit methods. The results across three datasets indicate that our approach improves performance with different OOD detection scores. Using the MCM score, AUROC and FPR95 improved by 2.46% and 8.29%, respectively. This improvement is attributed to our scaling factor α being set to 0.8, which, when using MCM, amplifies the difference between ID and OOD scores, thus enhancing our method's effectiveness.

The Choice of LLMs. To verify our method on different LLMs, we have considered several LLMs with varying parameter sizes, as shown in Table 4. We conduct experiments using various LLMs to comprehensively assess the effectiveness of descriptors generated by different LLMs. Specifically, we utilize LLaMA-3-8b, ChatGPT-4, and Claude 2 for descriptor generation. The average results across three datasets indicate that using different LLMs achieved better performance compared to the baseline MCM. Additionally, LLaMA-3-8b outperforms

Table 3. Results after integrating our method with various scoring functions as baselines. The ID datasets are ImageNet-10, ImageNet-20, and Waterbirds, with corresponding OOD datasets being ImageNet-20, ImageNet-10, and Spurious OOD, respectively. "Average" represents the mean performance across these three datasets, and "Improvement" indicates the enhancement relative to the baseline.

Methods	Average		Improvement	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓
MCM	96.70	11.12	/	/
MCM_{ours}	99.16	2.50	+2.46	-8.62
Energy	97.49	11.49	/	/
$\text{Energy}_{\text{ours}}$	97.83	8.92	+0.34	-2.57
MaxLogit	97.67	10.84	/	/
$\text{MaxLogit}_{\text{ours}}$	98.01	8.82	+0.34	-2.02

Claude2 and GPT-4.0 in both AUROC and FPR95 metrics, demonstrating the generalizability and robustness of our method. This can be attributed to the fact that LLaMA-3-8b excels at generating short, task-specific text, which benefits OOD detection by providing focused descriptions [50]. In contrast, GPT-4, while powerful in broader tasks, may produce more verbose responses that could introduce noise into similarity comparisons. Therefore, the performance differences between the two are likely due to factors such as the relevance and specificity of generated OOD class names, the precision of descriptive terms, and prompt interpretation.

Table 4. Impact of using different LLMs on results, consistent dataset settings as in Table 3. "A" represents the AUROC, "F" represents the FPR95.

Methods	Average		Improve	
	A↑	F↓	A↑	F↓
MCM	96.70	11.12	/	/
LLaMA-3-8b	99.16	2.50	+2.46	-8.62
Claude2	99.03	3.21	+2.33	-7.91
GPT-4.0	99.06	3.00	+2.36	-8.12

Table 5. Impact of number of OOD classes on results, consistent dataset settings as in Table 3. k denotes the number of selected OOD classes.

Number	Average		Improve	
	A↑	F↓	A↑	F↓
MCM	96.70	11.12	/	/
k=1	98.99	2.92	+2.29	-8.20
k=2	99.00	2.53	+2.30	-8.59
k=3	99.27	2.20	+2.57	-8.92
k=4	99.06	2.43	+2.36	-8.69
k=5	99.08	2.74	+2.38	-8.38

Fine-grained Textual Features. To verify the effectiveness of generating specific descriptive terms for OOD classes, we created various types of descriptive information to evaluate performance, as shown in Table 4. We conduct experiment where descriptors are simplified to only use generated hard OOD classes for inference, altering the template to "A photo of {Class_Name}" to assess the

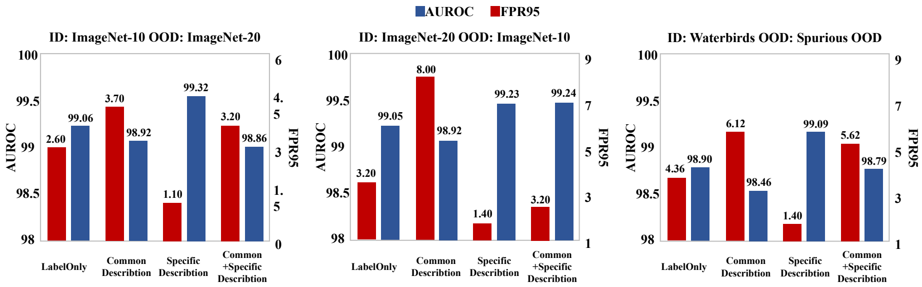


Fig. 4. Impact of prompt variation on LLM-generated descriptors.

efficacy of textual features. We refer to this as "Label only" in Fig.4. Additionally, to validate the effectiveness of our descriptors in distinguishing OOD from current ID samples, we designed different prompts for verification. Specifically, we modify the question in Fig.2 (b) to directly ask, "Please describe the visual characteristics of {OOD_CLASS}", "What are the visual features similar to {OOD_CLASS} and {ID_CLASS}?" to obtain descriptions that include both common and specific OOD features, as well as descriptions with only specific OOD features. In Fig.4, these are represented as "Common+Specific description" and "Common description", respectively. "Specific description" represents the primary method of this paper, obtaining unique features that distinguish OOD classes from ID classes.

The results indicate that even when only category names are used to provide textual information ("Label only"), our method still outperforms the baseline MCM. However, when the textual description includes a significant amount of ID features ("Common description"), the performance of OOD detection significantly decreases, with the most notable decline observed on the ImageNet-20 dataset, where FPR95 and AUROC drop by 4.80% and 0.52%, respectively. When the OOD descriptors include only the unique features that distinguish OOD classes from ID classes ("Specific description"), our method achieves the best performance across all three datasets. When descriptors include both types of features ("Common+Specific description"), the results on various datasets are better than those with only the common description, but still inferior to those with only the specific description. This validates that common features shared between OOD and ID classes are detrimental to OOD detection.

4.6 Hyperparameter Sensitivity Analysis

Number of OOD Class Labels. Investigating the impact of the number of generated OOD classes on performance, we set k with different values, i.e., {1, 2, 3, 4, 5}. As shown in Table 5, performance improves initially and then declines as k increases, with the best results at $k = 3$. The initial gain stems from the expanded textual space, enhancing the model's capacity to separate ID and OOD samples. However, larger k increases the likelihood of generating semantically

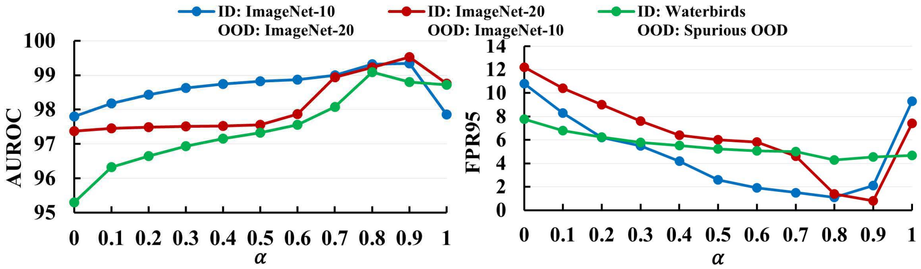


Fig. 5. Evaluation of hyperparameter α's effect on ID confidence correction, with FPR95 and AUROC trends in the left and right graphs.

similar OOD descriptions, causing feature overlap with ID or existing OOD classes. This overlap can be attributed to the inherent characteristics of LLMs, which may generate similar descriptions for semantically related concepts. For example, OOD descriptions for "cat" may include "kitty" and "kitten", blurring decision boundaries and impairing performance when $k > 3$.

ID Confidence Calibration. We investigate the influence of the ID confidence calibration coefficient on the performance, adjusting the rectification degree α across the range $\{0, 0.1, ..., 1\}$. We conduct experiments on the ImageNet-10, ImageNet-20, and Waterbirds datasets, and the results are shown in Fig.5. We observe that when α is set to extreme values of 0 and 1, and the performance of OOD detection significantly decreases. When α is 0, setting the confidence directly to 0 leads to some ID samples being incorrectly classified as OOD, reducing robustness. Conversely, when α is 1, not adjusting the confidence negates the model's effectiveness. However, when α is between 0.7 and 0.9, the model performs well across all three datasets, indicating that our model is not sensitive to the α parameter.

5 CONCLUSION

In this paper, we propose a simple yet effective zero-shot OOD detection approach that leverages LLMs to enhance textual feature extraction for both ID and OOD classes. Specifically, we design prompts to generate specific semantic text, integrating ID class information to improve OOD distinction. We calibrate ID confidence based on generated OOD scores and propose a regionally enhanced semantic OOD score for detection. Our approach guides VLMs to focus on relevant image regions, leading to significant performance gains. Extensive experiments show our method outperforms SOTA approaches across multiple benchmarks and VLM architectures.

Acknowledgments. This work is supported by the National Key RD Program of China (2022YFF0712100), NSFC (62276131), Natural Science Foundation of Jiangsu Province of China under Grant (BK20240081).

References

1. Abhijit Bendale, Terrance E. Boult: Towards Open World Recognition. In: CVPR (2015)
2. Dan Hendrycks, Kevin Gimpel: A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks, In: ICLR (2017)
3. Jingkang Yang, Kaiyang Zhou and Yixuan Li et al.: Generalized Out-of-Distribution Detection: A Survey, In: International Journal of Computer Vision (2024)
4. Jingyang Zhang, Jingkang Yang and Pengyun Wang et al.: OpenOOD v1.5: Enhanced Benchmark for Out-of-Distribution Detection, In: arXiv, abs/2306.09301 (2023)
5. Gao Huang, Zhuang Liu and Laurens van der Maaten et al., Densely Connected Convolutional Networks, In: CVPR (2017)
6. Di Feng, Ali Harakeh and Steven L. Waslander et al.: A Review and Comparative Study on Probabilistic Object Detection in Autonomous Driving, In: Trans. Intell. Transp. Syst (2022)
7. Li Chen, Penghao Wu and Kashyap Chitta et al.: End-to-end Autonomous Driving: Challenges and Frontiers, In: Transactions on Pattern Analysis and Machine Intelligence (2024)
8. Igor Kononenko: Machine learning for medical diagnosis: history, state of the art and perspective, In: Artif. Intell. Medicine (2001)
9. Yen-Chang Hsu, Yilin Shen and Hongxia Jin et al.: Generalized ODIN: Detecting Out-of-Distribution Image Without Learning From Out-of-Distribution Data, In: CVPR (2020)
10. Haoran Wang, Weitang Liu and Alex Bocchieri et al.: Can multi-label classification networks know what they don't know? In: NIPS (2021)
11. Vikash Sehwag, Mung Chiang and Prateek Mittal et al.: SSD: A Unified Framework for Self-Supervised Outlier Detection, In: ICLR (2021)
12. Kimin Lee, Kibok Lee and Honglak Lee et al.: A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In: NIPS (2018)
13. Weitang Liu, Xiaoyun Wang and John D. Owens et al.: Energy-based Out-of-distribution Detection. In: NIPS (2020)
14. Chentao Cao, Zhun Zhong and Zhanke Zhou et al.: Envisioning Outlier Exposure by Large Language Models for Out-of-Distribution Detection. In: ICML (2024)
15. Xue Jiang, Feng Liu and Zhen Fang et al.: Negative Label Guided OOD Detection with Pretrained Vision-Language Models. In: ICLR (2024)
16. Alec Radford, Jong Wook Kim and Chris Hallacy et al.: Learning Transferable Visual Models From Natural Language Supervision. In: ICML (2021)
17. Yifei Ming, Ziyang Cai and Jiuxiang Gu et al.: Delving into Out-of-Distribution Detection with Vision-Language Representations. In: NIPS (2022)
18. Hualiang Wang, Yi Li and Huifeng Yao et al.: CLIPN for Zero-Shot OOD Detection: Teaching CLIP to Say No. In: ICCV (2023)
19. Fellbaum, Christiane: WordNet: An electronic lexical database. In: MIT press (1998)
20. Yi Dai, Hao Lang and Kaisheng Zeng et al.: Exploring Large Language Models for Multi-Modal Out-of-Distribution Detection. In: EMNLP (2023)
21. Fabio Petroni, Tim Rockt́aschel and Sebastian Riedel et al.: Language Models as Knowledge Bases? In: EMNLP (2019)

22. Jihoon Tack, Sangwoo Mo and Jongheon Jeong et al.: CSI: Novelty Detection via Contrastive Learning on Distributionally Shifted Instances. In: NIPS (2020)
23. Rui Huang, Yixuan Li: MOS: Towards Scaling Out-of-Distribution Detection for Large Semantic Space. In: CVPR (2021)
24. Xuefeng Du, Xin Wang and Gabriel Gozum et al.: Unknown-Aware Object Detection: Learning What You Don't Know from Videos in the Wild. In: CVPR (2022)
25. Hongxin Wei, Renchunzi Xie and Hao Cheng et al.: Mitigating Neural Network Overconfidence with Logit Normalization. In: ICML (2022)
26. Yifei Ming, Yiyou Sun and Ousmane Dia et al.: CIDER: Exploiting Hyperspherical Embeddings for Out-of-Distribution Detection. In: arXiv, abs/2203.04450 (2022)
27. Xue Jiang, Feng Liu and Zhen Fang et al.: Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In: ICLR (2018)
28. Yang Yang, Yuxuan Zhang and Xin Song et al.: Not All Out-of-Distribution Data Are Harmful to Open-Set Active Learning. In: NIPS (2023)
29. Wenjuan Xi, Xin Song and Weili Guo et al.: Robust Semi-Supervised Learning for Self-learning Open-World Classes. In: ICDM (2023)
30. Yang Yang, Nan Jiang and Yi Xu et al.: Robust Semi-Supervised Learning by Wisely Leveraging Open-Set Data. In: Trans. Pattern Anal. Mach. Intell. (2024).
31. Yang Yang, Hongchen Wei and Zhen-Qiang Sun et al.: S2OSC: A Holistic Semi-Supervised Approach for Open Set Classification. In: ACM Trans. Knowl. Discov. Data. (2022).
32. Rui Huang, Andrew Geng and Yixuan Li et al.: On the Importance of Gradients for Detecting Distributional Shifts in the Wild. In: NIPS (2021)
33. Yiyou Sun, Yifei Ming and Xiaojin Zhu et al.: Out-of-Distribution Detection with Deep Nearest Neighbors. In: ICML (2022)
34. LeCun, Yann and Chopra et al.: A tutorial on energy-based learning. In: Predicting structured data (2006).
35. Haoqi Wang, Zhizhong Li and Litong Feng et al.: ViM: Out-Of-Distribution with Virtual-logit Matching. In: CVPR (2022)
36. Sepideh Esmaeilpour, Bing Liu and Eric Robertson et al.: Zero-Shot Out-of-Distribution Detection Based on the Pre-trained Model CLIP. In: AAAI (2022)
37. Tom B. Brown, Benjamin Mann and Nick Ryder et al.: Language Models are Few-Shot Learners. In: NIPS (2020)
38. Hugo Touvron, Louis Martin and Kevin Stone et al.: Llama 2: Open Foundation and Fine-Tuned Chat Models. In: arXiv, abs/2307.09288 (2023).
39. OpenAI: GPT-4 Technical Report. In: arXiv, abs/2303.08774 (2023).
40. Jia Deng, Wei Dong and Richard Socher et al.: ImageNet: A large-scale hierarchical image database. In: CVPR (2009)
41. Grant Van Horn, Oisin Mac Aodha and Yang Song et al.: The INaturalist Species Classification and Detection Dataset. In: CVPR (2018)
42. Jianxiong Xiao, James Hays and Krista A. Ehinger et al.: SUN database: Large-scale scene recognition from abbey to zoo. In: CVPR (2010)
43. Bolei Zhou, Àgata Lapedriza and Aditya Khosla et al.: Places: A 10 Million Image Database for Scene Recognition. In: Trans. Pattern Anal. Mach. Intell. (2018).
44. Mircea Cimpoi, Subhransu Maji and Iasonas Kokkinos et al.: Describing Textures in the Wild. In: CVPR (2014)
45. Shiori Sagawa, Pang Wei Koh and Tatsunori B. Hashimoto et al.: Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. In: arXiv, abs/1911.08731 (2019)
46. Yifei Ming, Hang Yin and Yixuan Li et al.: On the Impact of Spurious Correlation for Out-of-Distribution Detection. In: AAAI (2022)

47. Leitian Tao, Xuefeng Du and Jerry Zhu et al.: Non-parametric Outlier Synthesis. In: ICLR (2023)
48. Piyush Sharma, Nan Ding and Sebastian Goodman et al.: Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In: ACL (2018)
49. Yichen Bai, Zongbo Han and Bing Cao et al.: ID-like Prompt Learning for Few-Shot Out-of-Distribution Detection. In: CVPR (2024)
50. Zongxi Li, Xianming Li and Yuzhang Liu et al.: Label Supervised LLaMA Finetuning. In: arXiv, abs/2310.01208 (2023)
51. Haonan Xu and Yang Yang: ITP: Instance-Aware Test Pruning for Out-of-Distribution Detection. In: AAAI (2025)
52. Yang Yang and Haonan Xu: Strengthen Out-of-Distribution Detection Capability with Progressive Self-Knowledge Distillation. In: ICML (2025)