TAG: Text Prompt Augmentation for Zero-Shot Out-of-Distribution Detection

Xixi Liu and Christopher Zach

Chalmers University of Technology
Gothenburg, Sweden
{xixil, zach}@chalmers.se

Abstract. Out-of-distribution (OOD) detection has been extensively studied for the reliable deployment of deep-learning models. Despite great progress in this research direction, most works focus on discriminative classifiers and perform OOD detection based on single-modal representations that consist of either visual or textual features. Moreover, they rely on training with in-distribution (ID) data. The emergence of vision-language models allows to perform zero-shot OOD detection by leveraging multi-modal feature embeddings and therefore only rely on labels defining ID data. Several approaches have been devised but these either need a given OOD label set, which might deviate from real OOD data, or fine-tune CLIP, which potentially has to be done for different ID datasets. In this paper, we first adapt various OOD scores developed for discriminative classifiers to CLIP. Further, we propose an enhanced method named *TAG* based on Text prompt AuGmentation to amplify the separation between ID and OOD data, which is simple but effective, and can be applied on various score functions. Its performance is demonstrated on CIFAR-100 and large-scale ImageNet-1k OOD detection benchmarks. It consistently improves AUROC and FPR95 on CIFAR-100 across four commonly used architectures over four baseline OOD scores. The average AUROC and FPR95 improvements are 6.35% and 10.67%, respectively. The results for ImageNet-1k follow a similar, but less pronounced pattern. The code is available at: https://github.com/XixiLiu95/TAG.

Keywords: Vision-language models · Zero-shot out-of-distribution detection

1 Introduction

To guarantee the safe deployment of deep learning models in the “wild,” particularly for high-stake applications such as autonomous driving [11] and intelligent health care [38], it is unarguably critical for the models to learn what they do not know [29]. For instance, models should be able to flag inputs highly unlikely according to the training distribution and avoid unreliable predictions for such data. Specifically, models are expected to identify samples that exhibit covariate shift (change in the input distribution) or semantic shift (change in the label distribution) depending on the use case [49].

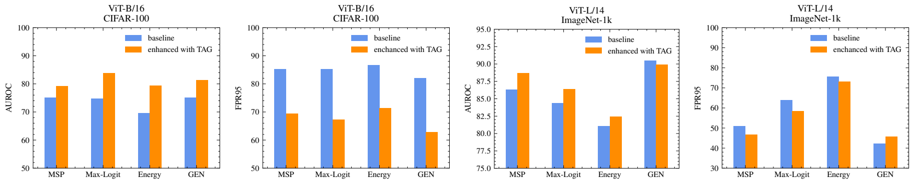

Fig. 1: *Effectiveness of TAG applied with 4 Baseline Scores on CIFAR100 (left 2 columns) and ImageNet-1k (right 2 columns).* Reported are AUROC values (%) and FPR95 (%). The averages are computed across 5 OOD datasets (CIFAR-100) and 4 OOD datasets (ImageNet-1k), respectively.

In this work, we focus on the case of identifying semantic shift. A plethora of the research in this setting uses standard discriminative classifiers [4, 7, 12–14, 17, 20, 22, 23, 39, 40, 43], where the label information is simply encapsulated as a one-hot vector. Therefore, the above-mentioned methods mostly rely on visual features extracted from the pre-trained models. Moreover, their OOD detection performance is highly correlated with the accuracy of the classifiers [23, 43]. The emergence of vision-language models (VLMs e.g. CLIP [34]) that learn the joint representation of image and text offers a great opportunity to exploit it for OOD detection, particularly, for the semantic shift. Specifically, the text prompt for each class, processed by the text encoder, can be viewed as a class prototype (in feature space). Unlike scenarios involving discriminative classifiers, where the models must undergo training on the ID dataset, CLIP-based methods only require the set of labels comprising the ID dataset. A number of works have observed that CLIP can be used as a powerful zero-shot OOD detector [5, 9, 10, 25, 28]. Following the definition in [9], zero-shot OOD detection means that only the names of the ID dataset can be utilized, and it does not access the training data of ID dataset. However, some of them either need to create the OOD label set manually [10] or generate the OOD label set automatically [9], which might diminish the OOD performance if the designed OOD label is not representative. MCM [25] is free from the pre-defined OOD labels but less effective in some cases. GL-MCM [28] enhances the performance of MCM by exploiting the local features, which to some extent restricts its deployment scenarios. [5] designs a pipeline for OOD detection by utilizing the external knowledge from large language models (LLMs) to generate descriptors for the ID dataset. In this work, we extend the score functions based on a single-modal regime (i.e., discriminative classifiers) to a multi-modal regime (i.e., CLIP) to perform zero-shot OOD detection. Furthermore, an enhanced method based on text prompt augmentation is proposed to further improve the performance. Figure. 1 highlights its performance compared to other baseline score functions.

Contribution We present a simple but effective method, *TAG* (for Text prompt AuGmentation), to enhance the performance of zero-shot OOD detection equipped with various score functions including MSP [13], MaxLogit [12], Energy [22], and GEN [23].

1. TAG only uses label information of the training data and is completely outlier-free (in terms of both OOD data and label information). It also does not require external knowledge from LLMs, sophisticated prompt ensembling or additional training, meaning it can be deployed in a wider range of scenarios.
2. It consistently achieves significantly better results under various score functions on CIFAR-100, and the improvement remains on ImageNet-1k across 4 architectures and 3 baseline OOD methods (Fig. 1 and Section 4).

2 Related Work

Vision-based OOD detection Performing OOD detection in terms of semantic shift on discriminative classifiers has been a long-standing research field [2, 4, 6–8, 12–14, 16, 17, 22, 23, 26, 27, 33, 40, 41, 46, 47], and can be roughly categorized based on whether the outliers are exposed during training. Firstly, the methods that do not require outlier exposure (OE) can be grouped into (i) deriving new score functions based on either logit information such as Energy [22] and MaxLogit [12], or predictive distribution such as MSP [13] and GEN [23]; additionally, GradNorm [17] utilizes the information from both features extracted from the penultimate layer and predictive distributions. (ii) utilizing the training feature statistics such as [20, 43] or the learned weight of the last fully connected layer [4] to devise OOD score. It is intuitive that using the information from the training data could further boost the performance of OOD detection. However, this is infeasible in the case when the training data is confidential or otherwise unavailable. (iii) enhancing the OOD performance by either obtaining distinct features to distinguish ID and OOD data such as ODIN [21], Generalized ODIN [16], ReAct [40], RankFeat [39], ASH [6], and SCALE [47], or augmenting softmax-based confidence scores with feature-agnostic information such as SIRC [46]. Those enhanced methods are compatible with several score functions including MSP [13], Energy [22], and GEN [23]. Additionally, unlike the training of a standard classifier using cross-entropy loss, [41] and CIDER [27] devise contrastive learning-based methods for OOD detection.

The methods required to access OOD data typically involve devising a new training loss with OE explicitly [14, 26] or implicitly [7, 8, 33, 45]. Specifically, [14] firstly propose to jointly optimize a classification loss and a regularization term that forces the predictive distribution of the OOD sample to be uniform. [26] proposes to perform outlier mining firstly by sampling a posterior distribution and then applying energy regularization [22] afterward. Additionally, [45] argues that the selected OOD data for training might deviate from the real OOD data and the performance of OE might degrade on the unseen OOD data. Therefore, a min-max learning scheme is formulated to search for the OOD samples that are most intriguing to the model and learn from such OOD data. However, heavier computation is required compared to other OE methods. [8] does not rely on any OOD data but instead obtains the OOD feature embeddings by sampling the low density of the training feature space. While [7] utilizes the learned text embeddings of the training data and draws samples from the low-density regime to obtain OOD text embeddings. Furthermore, the sampled OOD text embeddings are processed with Stable Diffusion [35] to generate synthetic OOD samples. Finally, energy regularization [22] is applied to enable the training for OOD detection. Nevertheless, implementing this method requires generating OOD data, in particular, for each ID dataset, thereby its applicability is restricted in various deployment scenarios.

Vision-language based OOD detection CLIP [34], as the most popular and publicly available VLM is getting recognition for the task of OOD detection [9, 10, 25, 44]. [10] is the first work to explore the capability of CLIP for zero-shot OOD detection. Specifically, two non-overlapped sets of label space including the class names of the ID dataset $\mathcal{Y}_{\text{ID}}$, and class names manually designed $\mathcal{Y}_{\text{OOD}}$ are created. During inference, an image embedding $\mathcal{I}(\boldsymbol{x})$ is obtained for each image $\boldsymbol{x}$, and applying Softmax to the logits $\boldsymbol{s}$ (i.e., the cosine similarity between the image embedding $\mathcal{I}(\boldsymbol{x})$ and all text embeddings), the predictive distribution is obtained and denoted by $\boldsymbol{p} = \text{Softmax}(\boldsymbol{s})$. Note $\boldsymbol{p}$ can be split to $p(\text{in}|\,\boldsymbol{x}) = \sum_{i \in \mathcal{Y}_{\text{ID}}} \boldsymbol{p}_i$ and $p(\text{out}|\,\boldsymbol{x}) = \sum_{i \in \mathcal{Y}_{\text{OOD}}} \boldsymbol{p}_i$, and $p(\text{in}|\,\boldsymbol{x}) + p(\text{out}|\,\boldsymbol{x}) = 1$. Finally, the OOD score is designed as $p(\text{in}|\,\boldsymbol{x}) = \sum_{i \in \mathcal{Y}_{\text{ID}}} \boldsymbol{p}_i$. To resolve the inconvenience of manually designed OOD labels arising from [10], ZOC [9] instead trains a text description generator to obtain $\mathcal{Y}_{\text{OOD}}$ automatically. First, a text-decoder denoted by $\text{Decoder}_{\text{text}}$ is trained on a large captioning data (i.e., a set of paired images and texts.). Afterward, the pre-trained $\text{Decoder}_{\text{text}}$ is used to generate an image description for each test image and then the top k words from the vocabulary with the highest probabilities are selected as $\mathcal{Y}_{\text{OOD}}$. The final label space is $\mathcal{Y}_{\text{ID}} \cup \mathcal{Y}_{\text{OOD}}$. The way to obtain the predictive distribution is the same as [10], but the final OOD score is defined as $1 - \sum_{i \in \mathcal{Y}_{\text{ID}}} \boldsymbol{p}_i$. Although [9, 10] demonstrated superior performance on OOD detection, they both rely on pre-defined OOD label sets, which unavoidably impedes their performance as the defined OOD labels might deviate from the real OOD label. Unsatisfactorily, the OOD label set potentially has to be designed for every ID dataset. Instead, CLIPN [44] fine-tunes the CLIP by introducing an additional text encoder on par with negative (learnable) prompts. The training loss incorporates two key components: image-text binary-opposite loss, which aims to align the image embedding with its unrelated negative text embedding, and the text semantic-opposite loss, designed to maximize the l_2 distance between two text embeddings with opposing meanings. The final OOD score is calculated either through the competing-to-win (CTW) algorithm or through the agreeing-to-differ (ATD) algorithm. However, the fine-tuning of CLIP inevitably has to be done for each ID dataset. MCM [25] instead neither depends on the design of the OOD label nor requires additional fine-tuning. It directly uses the text embeddings processed from the prompts `this is a photo of a` $\langle y_k \rangle$ as the concept prototypes to perform OOD detection. Our method TAG does not require both pre-defined OOD labels and pre-training. Moreover, it can be applied to MCM [25], potentially enhancing the performance of OOD detection.

Prompt engineering with external knowledge To improve the performance of zero-shot visual classification using VLMs, DCLIP [24] extends the default prompt for each class with its corresponding descriptions generated by LLMs (e.g., GPT-3). Instead, WaffleCLIP [36] empirically shows that replacing the generated GPT-3 descriptions with random word or character sequences leads to competitive performance. [5] explores to design a multi-modal OOD framework by utilizing the external knowledge from LLMs. However, additional calibration methods are required to maintain the quality of generated descriptors because of the hallucination of LLMs [1]. Different from [24, 36], our method is devised for the task of OOD detection and solely rely on the default prompt without any external knowledge. Moreover, our method can be integrated with DCLIP [24] and WaffleCLIP [36], potentially enhancing the performance of OOD detection.

3 Text Prompt Augmentation

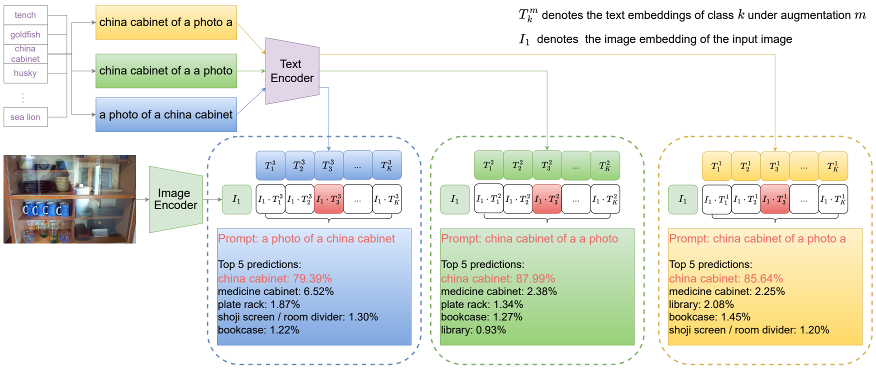
tench
goldfish
china cabinet
husky
sea lion
china cabinet of a photo a
china cabinet of a a photo
a photo of a china cabinet
Text Encoder
T_k^m denotes the text embeddings of class k under augmentation m
I_1 denotes the image embedding of the input image
Image Encoder
I_1
T_1^3 T_2^3 T_3^3 ... T_K^3
$I_1 \cdot T_1^3$ $I_1 \cdot T_2^3$ $I_1 \cdot T_3^3$ ... $I_1 \cdot T_K^3$
Prompt: a photo of a china cabinet
Top 5 predictions:
china cabinet: 79.39%
medicine cabinet: 6.52%
plate rack: 1.87%
shoji screen / room divider: 1.30%
bookcase: 1.22%
T_1^2 T_2^2 T_3^2 ... T_K^2
$I_1 \cdot T_1^2$ $I_1 \cdot T_2^2$ $I_1 \cdot T_3^2$ ... $I_1 \cdot T_K^2$
Prompt: china cabinet of a a photo
Top 5 predictions:
china cabinet: 87.99%
medicine cabinet: 2.38%
plate rack: 1.34%
bookcase: 1.27%
library: 0.93%
T_1^1 T_2^1 T_3^1 ... T_K^1
$I_1 \cdot T_1^1$ $I_1 \cdot T_2^1$ $I_1 \cdot T_3^1$ ... $I_1 \cdot T_K^1$
Prompt: china cabinet of a photo a
Top 5 predictions:
china cabinet: 85.64%
medicine cabinet: 2.25%
library: 2.08%
bookcase: 1.45%
shoji screen / room divider: 1.20%

Fig. 2: *Probabilities of Top-5 Predictions Using Different Sequences of the Default Text Prompt.* Class names are taken from ImageNet-1k and ViT-B/16 is used as backbone. The shuffled prompts of the non-target class are omitted for clean visualization.

CLIP is a vision-language model and consists of a text encoder $\mathcal{T}$ and an image encoder $\mathcal{I}$. Hundreds of millions of paired images and texts equipped with InfoNCE [32] loss are used for its training. To perform OOD detection using CLIP for a given ID dataset denoted by $\mathcal{D}_{\text{in}}$ with label space denoted by $\mathcal{Y}_{\text{in}} = \{y_1, y_2, \cdots, y_K\}$, the default text prototype t_k for the class k can be constructed as `a photo of` $\langle y_k \rangle$. During inference, a test image $\boldsymbol{x}$ is firstly processed by image encoder $\mathcal{I}$, we can re-interpret the cosine similarity s_k between extracted feature $\mathcal{I}(\boldsymbol{x})$ and all text prototypes $\mathcal{T}(t_k)$ as the logit, which is further normalized by Softmax, the probability that image $\boldsymbol{x}$ belong to class k can be calculated as

$$p_k(\boldsymbol{x}\,|\mathcal{Y}_{\text{in}}, \mathcal{I}, \mathcal{T}) = \frac{\exp(s_k/\tau)}{\sum_{j=1}^{K}\exp(s_j/\tau)}, \tag{1}$$

where $s_k = \frac{\mathcal{I}(\boldsymbol{x})\cdot\mathcal{T}(t_k)}{\|\mathcal{I}(\boldsymbol{x})\|\cdot\|\mathcal{T}(t_k)\|}$, and τ is a temperature parameter. By this interpretation, various score functions such as MSP [13], MaxLogit [12], Energy [22], and GEN [23] developed for the discriminative classifier can be applied to CLIP to perform OOD detection. The most significant benefit of using CLIP is that there is no need to access the training data of the ID dataset since the set of semantic labels for the ID dataset is the sole requirement.

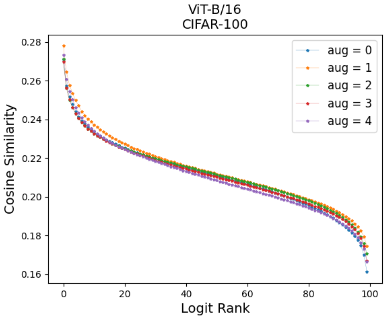

Fig. 3: *Average Sorted Cosine Similarity.* The dataset is CIFAR-100 [19], and $M = 5$ different text prompt augmentations are applied.

Sequence of the prompt The default text prompt for CLIP includes but is not limited to `a photo of a` $\langle y_k \rangle$. We observe that it is not necessary to use the right order of the text prompt. Instead, with a grammatically incorrect sequence $\langle y_k \rangle$ `of a a photo`, CLIP may still yield a correct classification, sometimes even with a higher probability for the target class. An example is illustrated in Figure 2. Here the default prompt is randomly shuffled and a classification task on ImageNet-1k [37] is performed based on the shuffled prompt. One can see that the image of the china cabinet is correctly classified with higher probability using the incorrect order of text prompt.

Effect of text prompt augmentation It is empirically observed that the cosine similarity is non-uniform for the ID dataset, which is also noticed by MCM [25]. Moreover, we also observe that this phenomenon consistently occurs when different text augmentations are applied. The average cosine similarity of CIFAR100 applied with 5 different random text augmentations is shown in Figure 3.

TAG Motivated by the aforementioned phenomenon, an enhanced method is proposed to improve the performance of OOD detection under various score functions. Specifically, M augmented text prompts for each class k can be obtained by randomly shuffling the default prompt that CLIP[1] uses. The PyTorch-like code for generating M different augmented tokens (i.e. tokenized text prompt) is presented in Algorithm. 1. Each augmented set is denoted by $t^m = \{t_1^m, t_2^m, \cdots, t_K^m\}$, where t_k^m denotes the augmented text prompt for class k under augmentation m. After obtaining M sets of text prompts, the probability that the test sample $\boldsymbol{x}$ belonging to class k with the text prompt augmentation t_k^m is calculated as

$$p_k^m(\boldsymbol{x}\,|\mathcal{Y}_{\text{in}}, \mathcal{I}, \mathcal{T}) = \frac{\exp(s_k^m/\tau)}{\sum_{j=1}^{K}\exp(s_j^m/\tau)}, \tag{2}$$

where $s_k^m = \frac{\mathcal{I}(\boldsymbol{x})\cdot\mathcal{T}(t_k^m)}{\|\mathcal{I}(\boldsymbol{x})\|\cdot\|\mathcal{T}(t_k^m)\|}$ is the logit of class k with text-prompt m, and τ is the temperature hyper-parameter. Assuming MSP [13] is used as the OOD score to perform OOD detection, meaning

$$S^m(\boldsymbol{x}) = \max_k \, p_k^m, \tag{3}$$

the final score function for OOD detection is

$$S(\boldsymbol{x}) = \tfrac{1}{M}\sum_{m=1}^{M} S^m(\boldsymbol{x}). \tag{4}$$

The alternative scoring methods including MaxLogit [12], Energy [22], and GEN [23] can also be utilized by substituting the Eq. 3 with the respective score functions.

Logits vs. probabilities In [25] it is argued that using the maximum probability (MSP/MCM) instead of the maximum logit (MaxLogit) is beneficial in terms of the FPR (Theorem 1 in [25]). In particular, for a sufficiently large choice of τ, MSP/MCM always yields a lower FPR than MaxLogit (under a certain assumption on the values of the non-maximal logits). In the supplementary material we improve on their result by replacing the specific assumption on the logits (Assumption A.1 in [25]) with a simple assumption that the logits are bounded from below. This assumption is clearly satisfied for logits obtained as the cosine similarity between embedding vectors as they are constrained to the range $[-1, 1]$ by construction. We also want to point out that these theoretical results should be understood with some caution as by increasing τ only the FPR is controlled but not the TPR. This implies that very large values for τ will eventually be detrimental for the TPR, and a universal advantage of MSP/MCM over MaxLogit is not established.

4 Experiments

All experiments are conducted on two OOD benchmarks including CIFAR-100 [19] and ImageNet-1k [37]. We closely follow the evaluation protocol con-

[1] `a photo of a` $\langle y_k \rangle$

Algorithm 1: Generation of augmented tokens

```
# M: number of augmentations applied to the text prompt
# dataset: the ID dataset
def ShufflePrompt(words, c):
    random.shuffle(words) #  Shuffle the words randomly
    shuffled = ’ ’.join(words) # Reconstruct the shuffled prompt
    shuffled = shuffled.replace("classname", c)
    return shuffled
# Ensure that multiple-word class names are not split after shuffling
prompt = “a photo of a classname”
words = prompt.split() # Tokenize the prompt into words
MShuffledToken= [ ]
for m in range(M):
    TokenShuffled = [ ]
    for c in dataset.classes:
        text = ShufflePrompt(words, c)
        TokenShuffled.append(clip.tokenize(text))
    AllToken = torch.cat(TokenShuffled)
    MShuffledToken.append(AllToken)
```

ducted in [7, 27] with the CIFAR-100 as the ID dataset. For ImageNet-1k [37], we follow the evaluation done by ViM [43] and GEN [23]. All pre-trained checkpoints of CLIP models including ViT-based and ResNet-based are provided by OpenAI[2].

Models CLIP is used to demonstrate the effectiveness of our method. We use 5 models released by CLIP, which can be grouped into 1) ViT-based models including ViT-B/16, ViT-B/32, and ViT-L/14, in which the vision transformer (ViT) is used as the image encoder. 2) ResNet-based models including ResNet-50 and ResNet-101, in which the ResNet is taken as the image encoder. The text encoders are either a Continuous Bag of Words (COBW) model or a text transformer.

Datasets We perform OOD detection on a small-scale dataset with CIFAR-100 [19] as the ID dataset and a more realistic large-scale dataset with ImageNet-1k as the ID dataset. While CIFAR-100 has fewer classes compared to ImageNet-1k, the objects in the images are commonly centered and apparent. However, the objects in ImageNet-1k are sometimes rather small and sometimes partially occluded. For CIFAR-100 as ID dataset, the corresponding five OOD datasets are SVHN [30], iSUN [48], Places365 [52], Textures [3], and LSUN [50]. For the ImageNet-1k [37] as ID dataset, four commonly-used challenging OOD datasets are employed including ImageNet-O [15], Open-Image-O [18], Textures [3], and iNaturalist [42].

[2] `https://github.com/openai/CLIP`

Table 1: *Per-Dataset Performance of OOD Detection Methods and the Ones Enhanced with TAG denoted with* $*$. The image encoders are ViT-L/14 and ResNet-101. The ID dataset is **CIFAR-100**. The number of augmentation $M = 10$ for TAG. The temperature $\tau = 0.01$ for all methods. Green indicates improvement and red indicates degradation.

OOD method	SVHN		iSUN		Places365		Textures		LSUN		Average	
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
ViT-L/14												
MSP [13]	90.54	51.91	84.23	75.26	65.52	95.99	72.11	91.37	83.03	75.71	79.09	78.05
MSP*	92.91	34.31	89.77	50.94	69.47	90.08	77.19	80.6	86.27	64.34	83.12 (↑4.03)	64.05 (↓14.0)
MaxLogit [12]	88.62	69.28	82.97	80.32	92.39	33.97	91.01	38.09	62.72	93.97	83.46	70.85
MaxLogit*	88.17	69.34	84.06	77.8	92.88	32.29	91.16	37.94	63.0	94.2	83.85 (↑0.39)	62.31 (↓8.54)
Energy [22]	81.68	89.97	86.01	72.68	90.13	45.25	82.39	66.13	59.72	95.11	79.99	73.83
Energy*	81.36	87.39	77.06	86.97	94.15	28.3	90.88	40.8	50.48	95.34	78.79 (↓1.20)	67.76 (↓6.07)
GEN [23]	94.69	30.55	86.2	74.58	60.77	99.33	67.15	97.46	83.51	79.1	78.46	76.20
GEN*	94.13	32.46	89.37	61.77	62.52	99.04	70.42	94.34	86.11	64.74	80.51 (↑2.05)	70.47 (↓5.73)
MCM [25]	93.25	45.23	86.15	77.22	62.58	98.57	69.57	96.22	84.12	79.55	79.13	79.36
MCM*	94.13	32.68	90.06	55.76	64.99	97.44	73.34	91.08	86.65	64.54	81.83(↑2.70)	68.30(↓11.06)
ResNet-101												
MSP [13]	93.12	34.72	71.32	88.07	44.25	99.16	63.26	92.98	81.1	68.21	70.61	76.63
MSP*	95.9	24.21	79.18	75.6	46.09	98.92	65.31	90.99	88.15	50.55	74.93 (↑4.32)	68.05 (↓8.58)
MaxLogit [12]	96.47	19.63	79.6	79.1	83.05	50.57	81.8	55.85	73.02	92.31	82.79	59.49
MaxLogit*	98.76	5.58	78.45	85.38	82.38	51.33	85.55	45.96	74.98	92.04	84.02 (↑1.23)	56.06 (↓3.43)
Energy [22]	89.9	56.88	76.44	85.98	88.95	38.98	82.98	57.87	60.29	96.44	79.71	67.23
Energy*	95.75	26.42	70.63	91.48	88.17	40.26	86.64	47.02	58.67	97.79	79.97 (↑0.26)	60.59 (↓6.64)
GEN [23]	98.17	9.8	71.5	89.41	39.66	99.99	59.47	98.42	83.09	69.36	70.38	73.40
GEN*	98.47	5.24	82.2	76.16	44.1	99.87	63.33	95.85	91.59	45.47	75.94 (↑5.56)	64.52 (↓8.88)
MCM [25]	96.13	25.33	72.41	90.17	41.08	99.83	61.81	96.72	83.11	69.36	70.91	76.28
MCM*	97.38	18.29	81.25	78.49	44.8	99.77	64.76	95.21	90.31	50.8	75.70(↑4.79)	68.51(↓7.77)

Score functions Several commonly-used score functions derived for discriminative classifiers including MSP [13], MaxLogit [12], Energy [22], and GEN [23] are selected as the baseline methods. As suggested by GEN [23], we use top 100 classes and set $\gamma = 0.1$. Moreover, the score function MCM [25] (i.e. MSP with $\tau = 1$) designed for multi-modal models is also selected as one of the baselines.

Evaluation metrics The area under the receiver operating characteristic curve (AUROC) and FPR95 — the false positive rate when the true positive rate is 95%- are commonly utilized for the evaluation of OOD detection. Higher values of AUROC indicate better performance and lower values of FPR95 are better. The reported units for both metrics in all tables are percentages.

4.1 OOD Detection Experimental Results

In this section, the results of OOD detection using four score functions devised for discriminative classifiers but adapted to CLIP are presented first. Additionally, the score function MCM [25] designed for CLIP is also presented. Furthermore, the results of OOD detection enhanced with TAG denoted with $*$ are reported for each baseline score function. The experiments are running on NVIDIA GeForce RTX 2080Ti, CUDA 11.2 + PyTorch 2.1.0.

Table 2: *Per-Dataset Performance of OOD Detection Methods and the Ones Enhanced with TAG denoted with* $*$. The image encoders are ViT-L/14 and ResNet-101. The ID dataset is **ImageNet-1k**. The number of augmentation $M = 10$ for TAG. The temperature $\tau = 0.01$ for all methods except for MCM [25]. Green indicates improvement and red indicates degradation.

OOD method	**OpenImage-O**		**Textures**		**iNaturalist**		**ImageNet-O**		**Average**	
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
ViT-L/14										
MSP [13]	89.85	41.76	83.13	59.11	91.52	37.14	80.74	65.65	86.31	50.92
MSP*	92.42	34.34	85.92	55.0	93.61	31.4	82.84	66.55	88.70 (↑2.39)	46.82 (↓4.10)
MaxLogit [12]	90.27	50.82	74.63	83.41	91.66	49.91	81.08	71.5	84.41	63.91
MaxLogit*	91.54	44.17	80.05	74.38	92.57	43.91	81.44	70.8	86.40 (↑1.99)	58.32 (↓5.59)
Energy [22]	87.31	67.56	69.63	89.63	88.48	68.57	78.89	76.8	81.08	75.64
Energy*	87.64	65.17	74.85	83.84	88.72	65.56	78.63	77.7	82.46 (↑1.38)	73.07 (↓2.57)
GEN [23]	93.96	29.97	87.48	53.59	95.76	22.77	84.75	62.95	90.49	42.32
GEN*	93.72	31.68	86.85	55.85	94.79	28.37	84.33	67.35	89.92 (↓0.57)	45.81 (↑3.49)
MCM [25]	93.08	35.04	86.62	55.66	94.96	28.3	82.59	68.55	89.31	46.89
MCM*	93.05	36.96	88.55	52.02	93.9	37.22	82.04	73.8	89.39 (↑0.08)	50.00 (↑3.11)
ResNet-101										
MSP [13]	83.53	60.68	79.36	66.94	82.35	61.86	70.47	82.4	78.93	67.97
MSP*	85.39	59.03	82.62	61.24	85.61	58.88	71.72	84.4	81.34 (↑2.41)	65.89 (↓2.08)
MaxLogit [12]	83.94	72.86	69.61	91.96	82.33	82.9	71.78	86.05	76.91	83.44
MaxLogit*	84.69	72.62	75.47	88.53	83.24	79.85	72.08	86.7	78.87 (↑1.96)	81.92 (↓1.52)
Energy [22]	79.56	85.26	62.19	97.23	77.53	94.16	69.36	87.75	72.16	91.10
Energy*	79.2	84.13	67.19	95.27	77.11	92.32	69.15	89.35	73.16 (↑1.00)	90.27 (↓0.83)
GEN [23]	89.24	52.86	84.99	62.46	89.58	53.15	77.23	82.25	85.26	62.68
GEN*	88.48	55.89	85.01	65.33	89.12	57.53	76.31	84.15	84.73 (↓0.53)	65.72 (↑3.04)
MCM [25]	88.82	54.82	86.26	59.28	89.93	53.35	75.15	83.6	85.04	62.76
MCM*	88.38	56.27	88.25	51.53	89.21	57.12	75.36	84.3	85.30 (↑0.26)	62.31 (↓0.45)

Results on CLIP-ViT-L/14 and CLIP-ResNet-101 Two OOD benchmarks are selected to perform OOD detection. The results of CIFAR-100 are shown in Table. 1. First, the first block in Table. 1 indicates that our method (TAG) consistently and significantly improves the performance of OOD detection under five different scores in terms of FPR95. Moreover, the performance gain is also present for ResNet-101 by looking at the second block of Table 1. Particularly, MaxLogit [12] enhanced by TAG achieves the highest AUROC values and lowest FPR95 values on both ViT-L/14 and ResNet-101. The results of ImageNet-1k are shown in Table. 2. One can see that TAG again consistently improves the performance when using MSP [13], MaxLogit [12], and Energy [22] in terms of both AUROC and FPR95. When using GEN [23] as the OOD score, TAG is less effective on ImageNet-1k compared to CIFAR-100. We think this might be attributed to the limited capacity of pre-trained CLIP models. Specifically, the text prompt used in the training of CLIP is less informative, i.e., `a photo of` $\langle y_k \rangle$, where $\langle y_k \rangle$ is a noun and there is no other information such as activity information (i.e. verb) is provided. Moreover, the label information itself is quite restricted since there might be more than one object in the image [51].

Table 3: *Averaged Performance of Various OOD Detection Methods and the Ones Enhanced by TAG denoted with* $*$. Results are shown for ViT-B/16, ViT-B/32, and ResNet-50. For CIFAR-100, averages are computed across 5 OOD datasets, while for ImageNet-1k, the averages are derived from 4 OOD datasets. Green indicates improvement and red indicates degradation.

	OOD Method	**ViT-B/16**		**ViT-B/32**		**ResNet-50**		**Average**	
		AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
CIFAR-100	MSP [13]	75.05	85.30	77.05	75.45	60.25	90.01	70.78	83.59
	MSP*	79.21	69.48	78.45	74.20	71.55	62.55	76.40 (↑5.62)	68.74 (↓14.85)
	MaxLogit [12]	74.70	85.30	83.56	66.59	50.74	92.22	69.67	81.37
	MaxLogit*	83.86	67.32	87.03	59.84	75.33	78.57	82.07 (↑12.40)	68.58 (↓12.79)
	Energy [22]	69.64	86.62	80.05	71.51	45.27	93.20	64.99	83.78
	Energy*	79.31	71.45	84.07	64.56	69.81	85.75	77.73 (↑12.74)	73.92 (↓9.86)
	GEN [23]	75.07	82.14	77.98	66.36	60.38	84.90	71.14	77.80
	GEN*	81.38	62.81	78.23	68.31	71.89	63.28	77.17 (↑6.03)	64.80 (↓13.00)
	MCM [25]	75.55	84.76	77.93	73.00	60.12	88.30	71.2	82.02
	MCM*	80.85	65.75	78.62	75.06	71.93	63.33	77.13 (↑5.93)	68.05 (↓13.97)
ImageNet-1k	MSP [13]	82.85	59.36	79.79	65.00	79.22	67.41	80.62	63.92
	MSP*	85.13	57.76	82.03	64.63	81.10	65.33	82.75 (↑2.13)	62.57 (↓1.35)
	MaxLogit [12]	82.84	68.00	80.03	72.35	78.34	80.11	80.40	73.49
	MaxLogit*	84.48	65.92	82.54	67.98	79.09	79.90	82.03 (↑1.63)	71.26 (↓2.23)
	Energy [22]	79.26	79.09	76.48	82.03	74.11	88.66	76.61	83.26
	Energy*	80.23	79.92	78.73	78.48	74.16	88.73	77.71 (↑1.10)	82.38 (↓0.88)
	GEN [23]	88.70	50.09	86.64	56.34	86.02	59.19	87.12	55.21
	GEN*	87.83	54.95	85.64	62.65	84.46	65.53	85.98 (↓1.14)	61.04 (↑5.83)
	MCM [25]	88.18	51.9	86.31	55.45	86.09	57.17	86.86	54.83
	MCM*	87.72	56.94	86.31	59.08	85.54	62.02	86.52 (↓0.34)	59.34 (↑4.51)

Averaged results on other architectures To further investigate the effectiveness and robustness of TAG, we conducted OOD detection on three more models including two ViT-based models, which are ViT-B/16 and ViT-B/32, and one more ResNet-based model, ResNet-50. The performance is evaluated on both CIFAR-100 and ImageNet-1k. The results of CIFAR-100 are averaged over 5 different OOD datasets and shown in the top half of the Table. 3. It is undoubted that TAG again substantially and constantly improves the performance of all baseline score functions across 5 datasets and 3 architectures on CIFAR-100. Specifically, one can see that MaxLogit [12] enhanced by TAG achieves the best performance in terms of AUROC on average and GEN [23] enhanced by TAG obtains the lowest FPR95 values. For ImageNet-1k, the averages are calculated with 4 OOD datasets and shown in the bottom half of the Table. 3. TAG continually boosts the performance of OOD detection using MSP [13], MaxLogit [12] and Energy [22]. Additionally, the score function GEN [23] devised for the discriminative classifier achieves the best AUROC values and MCM [25] obtains the smallest FPR95 values on ImageNet-1k. In short, applying TAG on top of different score functions generally is a good idea to boost the performance fo OOD detection. Detailed results for each architecture can be found in supplementary material.

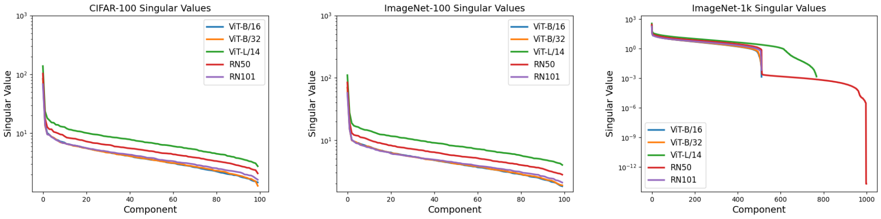

Fig. 4: The visualization of singular values for CIFAR-100, ImageNet-100, and ImageNet-1k.

4.2 Ablation studies

Analysis of text embeddings We observe that the improvement on ImageNet-1k is less pronounced than CIFAR-100. The hypothesis is that the pre-trained text embeddings for each class are not separable well. We confirm this by computing the rank of concatenated text embeddings and the visualization of singular values for CIFAR-100, ImageNet-100 and ImageNet-1k is shown in Fig. 4. One can see that the rank is 100 for both CIFAR-100 and ImageNet-100 across 5 different models. While the rank of the concatenated text embeddings for ImageNet-1k is generally less than 710 and most singular values are quite small. Detailed rank information with different models can be found in the supplementary material. We suspect that this is due to our utilized text prompts not covering the entire semantic space. Therefore we perform OOD detection on ImageNet-100, which is a subset of ImageNet-1k with 100 classes and the data list is provided by MCM [25]. The corresponding results can be found in Table. 4, and it is apparent that TAG consistently improves the baseline methods. MCM [25] combined with TAG is leading in terms of both AUROC and FPR95.

Table 4: *Per-Dataset Performance of OOD Detection Methods and the Ones Enhanced with TAG denoted with *.* The image encoders are ViT-L/14. The ID dataset is **ImageNet-100**. Green indicates improvement and red indicates degradation.

OOD Method	**OpenImage-O**		**Texture**		**iNaturalist**		**ImageNet-O**		**Average**	
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
MSP [13]	94.12	34.34	90.69	50.89	95.52	29.3	90.01	50.2	92.58	41.18
MSP*	95.55	26.41	92.65	43.97	96.31	21.74	91.35	46.3	93.97 (↑1.39)	34.60 (↓6.58)
MaxLogit [12]	93.97	38.39	83.83	75.14	94.95	34.93	90.58	51.95	90.83	50.10
MaxLogit*	94.99	30.29	88.79	58.39	95.79	25.28	91.31	45.6	92.72 (↑1.89)	39.89 (↓10.21)
Energy [22]	92.55	48.74	81.14	80.23	93.5	44.9	89.5	56.4	89.17	57.57
Energy*	93.11	43.28	86.12	66.59	94.17	36.71	89.86	51.85	90.82 (↑1.65)	49.61 (↓7.96)
GEN [23]	95.21	30.75	91.11	49.96	96.47	23.94	90.58	54.65	93.34	39.83
GEN*	95.3	31.46	94.02	37.34	95.81	30.51	90.64	54.7	93.94 (↑0.60)	38.50 (↓1.33)
MCM [25]	95.36	30.58	91.4	50.06	96.6	23.92	90.87	52.75	93.56	39.33
MCM*	95.64	28.22	94.06	38.39	96.2	26.17	91.1	51.45	94.25 (↑0.69)	36.06 (↓3.27)

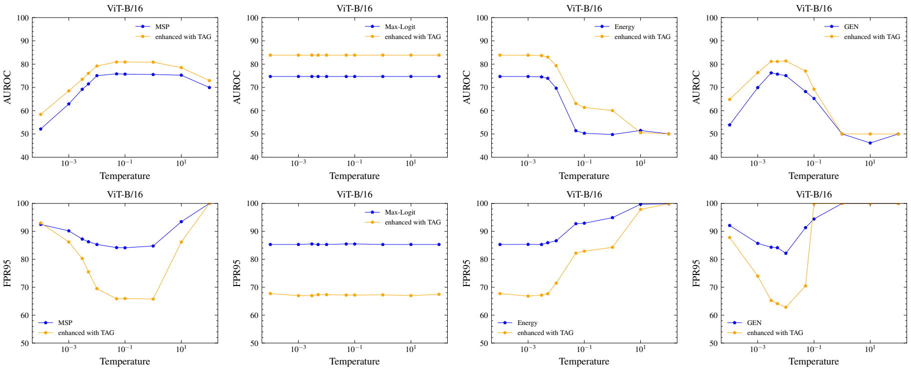

Fig. 5: *Averaged Performance (over 5 OOD Datasets) of TAG Applied with Different Temperature τ.* TAG performance in terms of AUROC values (top row) and FPR95 (bottom row). Each column denotes different score functions including MSP [13], MaxLogit [12], Energy [22], and GEN [23] (from left to right).

Choice of τ and M We empirically show the performance gap between the baseline methods and the ones enhanced with TAG using different temperatures τ and the number of text prompt augmentations M in terms of both AUROC and FPR95. Experiments of using different τ with CIFAR-100 as the ID dataset are conducted on ViT-B/16 and are presented in Figure. 5, in which each column represents one score function. The first row represents the results of regarding AUROC, and the second row indicates FPR95 performance. It is shown in Figure. 5 that TAG (with $M = 10$) could persistently improve the performance of the baseline OOD score in terms both of AUROC and FPR95 except for GEN [23] with $\tau = 0.1$. The evaluation regarding temperature τ for other architecture can be found in the supplementary material.

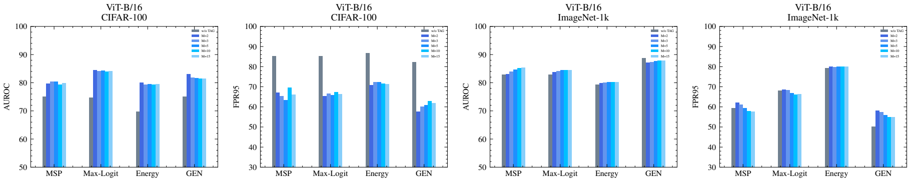

Fig. 6: *Averaged Performance of TAG Varying with Different Augmentations M.* The left two column corresponds to CIFAR-100 [19] dataset, and the right two columns corresponds to ImageNet-1k [37].

Additionally, we also investigate the effect of using different numbers of text prompt augmentations, and the results (with $\tau = 0.01$) on CIFAR-100 and ImageNet-1k are presented in Figure. 6. One can see that it is adequate to set $M = 2$ for CIFAR-100 as the ID dataset and $M = 10$ for ImageNet-1k as the

ID dataset. Results on other architectures can be found in the supplementary material.

Combining with DCLIP [24] and WaffleCLIP [36] We combine TAG with the default text prompt extended with descriptors generated by GPT-3 denoted by DCLIP [24] and prolonged with random characters or words denoted by WaffleCLIP [36]. The generated descriptors for each class are provided by WaffleCLIP [36]. CLIP means the default prompt `a photo of a` $\langle y_k \rangle$ is utilized. The OOD score is MSP with $\tau = 0.01$. One can see that TAG could further enhance the performance of OOD detection under various descriptors. WaffleCLIP [36] enhanced by TAG is leading in terms of AUROC. Results on other architectures with different score functions can be found in the supplementary material.

Table 5: *Performance of using different descriptors with $M = 10$ and $\tau = 0.01$.* The architecture is ViT-B/16. The ID dataset is ImageNet-1k [37]. * denotes the methods enhanced by TAG. The score function is MSP. Green indicates improvement and red indicates degradation.

Prompt	**OpenImage-O**		**Textures**		**iNaturalist**		**ImageNet-O**		**Average**	
	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
ViT-B/16										
CLIP [34]	86.39	51.51	81.57	61.12	87.53	49.66	75.92	75.15	82.85	59.36
CLIP*	89.18	45.71	84.13	60.04	89.69	48.09	77.52	77.2	85.13 (↑2.28)	57.76 (↓1.60)
DCLIP [24]	81.87	62.53	78.05	69.09	80.68	61.96	72.68	79.15	78.32	68.18
DCLIP*	86.3	57.91	83.26	63.84	84.4	70.26	75.4	81.7	82.34 (↑4.02)	68.43 (↑0.25)
WaffleCLIP [36]	83.19	59.88	79.89	67.42	82.49	61.04	75.0	75.9	80.14	66.06
WaffleCLIP*	88.72	47.78	85.63	55.48	87.46	55.67	78.82	74.4	85.16 (↑5.02)	58.33 (↓7.73)

5 Conclusion and Discussions

In this work we explore the benefits of adapting OOD scores designed for discriminative classifiers (e.g. trained with the cross-entropy loss) to vision-language models (i.e. CLIP trained with an InfoNCE [32] loss). Models like CLIP enable the use of various OOD scores to perform zero-shot OOD detection by only accessing the label information of the ID dataset, and they also allow variability in the resulting OOD scores by varying the text prompts. Our proposed method named TAG (Text prompt AuGmentation) leverages this variability, is easy to implement and effective for various OOD scores across different architectures with the minimal knowledge. It does not rely on the external knowledge from LLMs with the risk of hallucination or prompt ensembling. TAG offers significant improvements on standard OOD scores for most tested network models and datasets. A focus of future work is the less pronounced improvement on ImageNet-1k, which is likely to be attributed to the (simple) text prompts not exhausting CLIP's latent space, but may also be related to intrinsic shortcomings of the InfoNCE loss [31].

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

1. Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q.V., Xu, Y., Fung, P.: A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In: International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) (2023)
2. Bibas, K., Feder, M., Hassner, T.: Single layer predictive normalized maximum likelihood for out-of-distribution detection. In: NeurIPS (2021)
3. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: CVPR (2014)
4. Cook, M., Zare, A., Gader, P.: Outlier detection through null space analysis of neural networks. arXiv preprint arXiv:2007.01263 (2020)
5. Dai, Y., Lang, H., Zeng, K., Huang, F., Li, Y.: Exploring large language models for multi-modal out-of-distribution detection. In: Findings of the Association for Computational Linguistics: EMNLP 2023 (2023)
6. Djurisic, A., Bozanic, N., Ashok, A., Liu, R.: Extremely simple activation shaping for out-of-distribution detection (2023)
7. Du, X., Sun, Y., Zhu, X., Li, Y.: Dream the impossible: Outlier imagination with diffusion models. In: NeurIPS (2023)
8. Du, X., Wang, Z., Cai, M., Li, S.: Vos: Learning what you don't know by virtual outlier synthesis. In: ICLR (2022)
9. Esmaeilpourcharandabi, S., Liu, B., Robertson, E., Shu, L.: Zero-shot open set detection by extending clip. In: AAAI (2021)
10. Fort, S., Ren, J., Lakshminarayanan, B.: Exploring the limits of out-of-distribution detection. In: NeurIPS (2021)
11. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: CVPR (2012)
12. Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings. In: ICML (2022)
13. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: ICLR (2017)
14. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. In: ICLR (2019)
15. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. CVPR (2021)
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: CVPR (2020)
17. Huang, R., Geng, A., Li, Y.: On the importance of gradients for detecting distributional shifts in the wild. In: NeurIPS (2021)

18. Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Veit, A., Belongie, S., Gomes, V., Gupta, A., Sun, C., Chechik, G., Cai, D., Feng, Z., Narayanan, D., Murphy, K.: Openimages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://github.com/openimages (2017)
19. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. Master's thesis, Department of Computer Science, University of Toronto (2009)
20. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: NeurIPS (2018)
21. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: ICLR (2018)
22. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. In: NeurIPS (2020)
23. Liu, X., Lochman, Y., Zach, C.: Gen: Pushing the limits of softmax-based out-of-distribution detection. In: CVPR (2023)
24. Menon, S., Vondrick, C.: Visual classification via description from large language models. ICLR (2023)
25. Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., Li, Y.: Delving into out-of-distribution detection with vision-language representations. In: NeurIPS (2022)
26. Ming, Y., Fan, Y., Li, Y.: Poem: Out-of-distribution detection with posterior sampling. In: International Conference on Machine Learning (2022)
27. Ming, Y., Sun, Y., Dia, O., Li, Y.: Cider: Exploiting hyperspherical embeddings for out-of-distribution detection. arXiv preprint arXiv:2203.04450 (2022)
28. Miyai, A., Yu, Q., Irie, G., Aizawa, K.: Zero-shot in-distribution detection in multi-object settings using vision-language foundation models. In: AAAI (2023)
29. Nalisnick, E., Matsukawa, A., Teh, Y.W., Gorur, D., Lakshminarayanan, B.: Do deep generative models know what they don't know? In: ICLR (2019)
30. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.: Reading digits in natural images with unsupervised feature learning. In: NeurIPS (2011)
31. Oh, C., So, J., Byun, H., Lim, Y., Shin, M., Jeon, J.J., Song, K.: Geodesic multi-modal mixup for robust fine-tuning. In: NeurIPS (2023)
32. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. In: NeurIPS (2019)
33. Papadopoulos, A.A., Rajati, M.R., Shaikh, N., Wang, J.: Outlier exposure with confidence control for out-of-distribution detection. Neurocomputing
34. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2021)
36. Roth, K., Kim, J.M., Koepke, A.S., Vinyals, O., Schmid, C., Akata, Z.: Waffling around for performance: Visual classification with random words and broad concepts. In: ICCV (2023)
37. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. IJCV (2015)
38. Schlegl, T., Seeböck, P., Waldstein, S.M., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: IPMI (2017)
39. Song, Y., Sebe, N., Wang, W.: Rankfeat: Rank-1 feature removal for out-of-distribution detection. In: NeurIPS (2022)

40. Sun, Y., Guo, C., Li, Y.: React: Out-of-distribution detection with rectified activations. In: NeurIPS (2021)
41. Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-distribution detection with deep nearest neighbors. In: ICML (2022)
42. 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: CVPR (2018)
43. Wang, H., Li, Z., Feng, L., Zhang, W.: Vim: Out-of-distribution with virtual-logit matching. In: CVPR (2022)
44. Wang, H., Li, Y., Yao, H., Li, X.: Clipn for zero-shot ood detection: Teaching clip to say no. In: ICCV (2023)
45. Wang, Q., Ye, J., Liu, F., Dai, Q., Kalander, M., Liu, T., Hao, J., Han, B.: Out-of-distribution detection with implicit outlier transformation. In: ICLR (2023)
46. Xia, G., Bouganis, C.S.: Augmenting softmax information for selective classification with out-of-distribution data. In: ACCV (2022)
47. Xu, K., Chen, R., Franchi, G., Yao, A.: Scaling for training time and post-hoc out-of-distribution detection enhancement. In: ICLR (2024)
48. Xu, P., Ehinger, K.A., Zhang, Y., Finkelstein, A., Kulkarni, S.R., Xiao, J.: Turkergaze: Crowdsourcing saliency with webcam based eye tracking. arXiv preprint arXiv:1504.06755 (2015)
49. Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized out-of-distribution detection: A survey. arXiv preprint arXiv:2110.11334 (2021)
50. 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 (2016)
51. Yun, S., Oh, S.J., Heo, B., Han, D., Choe, J., Chun, S.: Re-labeling imagenet: from single to multi-labels, from global to localized labels. In: CVPR (2021)
52. 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 (2017)