

CLIPN for Zero-Shot OOD Detection: Teaching CLIP to Say No

Hualiang Wang, Yi Li, Huifeng Yao, Xiaomeng Li*

Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology

{hwangfd, ylini, yhfpro, eexmli}@ust.hk

Abstract

Out-of-distribution (OOD) detection refers to training the model on an in-distribution (ID) dataset to classify whether the input images come from unknown classes. Considerable effort has been invested in designing various OOD detection methods based on either convolutional neural networks or transformers. However, zero-shot OOD detection methods driven by CLIP, which only require class names for ID, have received less attention. This paper presents a novel method, namely CLIP saying “no” (CLIPN), which empowers the logic of saying “no” within CLIP. Our key motivation is to equip CLIP with the capability of distinguishing OOD and ID samples using positive-semantic prompts and negation-semantic prompts. Specifically, we design a novel learnable “no” prompt and a “no” text encoder to capture negation semantics within images. Subsequently, we introduce two loss functions: the image-text binary-opposite loss and the text semantic-opposite loss, which we use to teach CLIPN to associate images with “no” prompts, thereby enabling it to identify unknown samples. Furthermore, we propose two threshold-free inference algorithms to perform OOD detection by utilizing negation semantics from “no” prompts and the text encoder. Experimental results on 9 benchmark datasets (3 ID datasets and 6 OOD datasets) for the OOD detection task demonstrate that CLIPN, based on ViT-B-16, outperforms 7 well-used algorithms by at least 2.34% and 11.64% in terms of AUROC and FPR95 for zero-shot OOD detection on ImageNet-1K. Our CLIPN can serve as a solid foundation for effectively leveraging CLIP in downstream OOD tasks. The code is available on <https://github.com/xmed-lab/CLIPN>.

1. Introduction

Deep learning models [12, 6] have demonstrated excellent versatility and performance when the classes in the training and test datasets remain the same [32]. This is facilitated by the fact that the models are trained under com-

pletely closed-world conditions, meaning all encountered classes are in-distribution (ID) ones [7]. Nevertheless, these models tend to suffer from poor generalization and undesirable performance when deployed in real-world applications. This frustrating phenomenon is partially attributed to the existence of an enormous number of *unknown classes* distributed in the real world, which is challenging to detect as they have not been explicitly seen during the training stage.

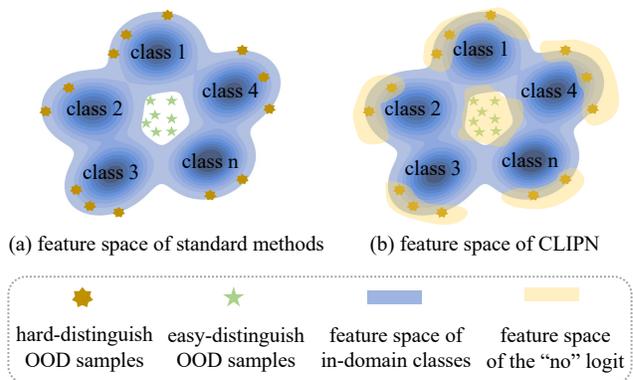


Figure 1. A toy comparison illustration of feature spaces between standard OOD detection algorithms and the proposed CLIPN. Our method involves a “no” logic, which provides a new feature space (yellow region) to directly identify OOD samples. The qualitative experiment visualization is shown in Figure. 5.

Out-of-distribution (OOD) detection task [2, 3, 9] has been raised and promptly attracted considerable interest from researchers. Briefly, the OOD detection task aims to empower the model to distinguish if the input images come from unknown classes. One of the mainstream OOD detection methods is to learn ID-specific features and classifiers, then develop the scoring function [42, 14] to metric how closely the input data matches the ID classes, which is measured by ID-ness [42] (or OOD-ness, an opposite case). For instance, MSP [14], MaxLogit [13], energy-based [27] and gradient-based [25] have been extensively employed to measure the ID-ness. Better ID-ness brings better OOD results. In summary, the key idea of these methods is to *teach ID knowledge to the model and then detect the dis-matched*

*Corresponding author.

cases referring to the model’s reply (score). The effectiveness of the above methods is seriously compromised by the following cases. As illustrated in Fig. 1, the green stars represent some OOD samples that are easy to distinguish, as they are relatively distant from all ID classes and naturally have high entropy, uniform probability [14], low logit [13] or low energy [27]. Conversely, hard-to-distinguish OOD samples (brown stars in Fig. 1) are more common and challenging. These samples are located relatively close to a certain ID class while being far away from other classes, resulting in high ID-ness. Therefore, existing methods such as those mentioned above fail to identify such samples accurately. As results shown in Fig. 5, even when we apply MSP [14] with different thresholding, there are still numerous mis-classified OOD samples, which are located in close proximity to ID classes.

Recently, some methods have sought to address the issue of hard-to-distinguish OOD samples by leveraging generalizable representations learned by CLIP [10], an open-world language-vision model trained on datasets with enormous volumes, such as Laion-2B [36]. Naturally, this task extends to zero-shot OOD detection (ZS OOD detection) [8, 30], which employs language-vision models to detect OOD data without requiring training on the ID dataset. ZOC [8] uses an additional text encoder to generate some candidate OOD classes not included in ID classes. Unfortunately, it is inflexible and unreliable when faced with a dataset containing a large number of ID classes, rendering it challenging to scale for large datasets such as ImageNet-1K [20]. MCM [30] leverages the text encoder component of CLIP and ID class prompts to obtain a more representative and informative ID classifier, which in turn enhances the accuracy of ID-ness estimates. However, this method still neglects to address the challenge of dealing with hard-to-distinguish OOD samples and suffers from limited performance; see results in Table 1.

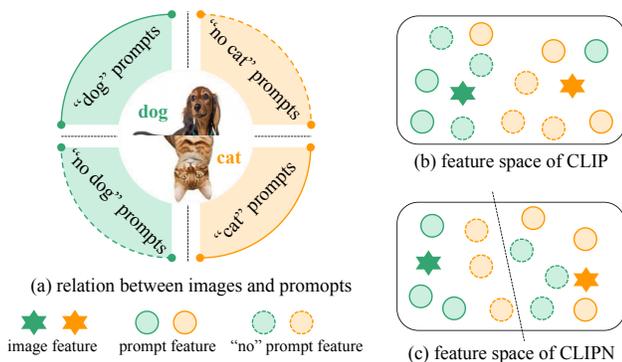


Figure 2. A toy illustration to determine that the original CLIP lacks “no” logic. The qualitative visualization is in Figure. 6.

Different from ZOC [8] and MCM [30], we attempt to exploit the *open-world* knowledge in CLIP to straightly

identify some hard-to-distinguish OOD samples even if their ID-ness is high. As the toy example shown in Figure. 2 (a), given a dog image and a cat image, we design four groups of prompts. Two groups contain class prompts with/of/.../having the photos of the dog or cat, while the other two groups use “no” prompts: a photo without/not of/.../not having the dog or cat. We conducted an experiment on CLIP to match the images with four prompts. Unfortunately, the results show that CLIP fails to accurately match the images, implying that it lacks “no” logic; as illustrated in the toy visualization in Fig. 2 (b) and qualitative visualization in Fig. 6.

To empower “no” logic within CLIP, we propose a new CLIP architecture, called CLIP saying “no” (CLIPN). It upgrades CLIP in terms of OOD detection in three ways. **(1) Architecture.** New “no” prompts and a “no” text encoder are added to CLIP. Our novel learnable “no” prompts integrate negation semantics within prompts, complementing the original CLIP’s prompts. Moreover, our “no” text encoder captures the corresponding negation semantics of images, making the CLIP saying “no” possible. **(2) Training Loss.** We further propose two loss functions. The first is image-text binary-opposite loss, which makes an image feature match with correct “no” prompt features. In other words, it can teach CLIP when to say “no”. The second is text semantic-opposite loss which makes the standard prompt and “no” prompts be embedded far away from each other. In other words, it can teach CLIP to understand the meaning of “no” prompts. **(3) Threshold-free Inference Algorithms.** After the training of CLIPN, we design two threshold-free algorithms: competing-to-win and agreeing-to-differ. The goal of competing-to-win is to select the most confident probability from standard and “no” text encoders as the final prediction. While agreeing-to-differ generates an additional probability for the OOD class by considering predictions from both standard and “no” text encoders. Experimental results on 9 benchmark datasets (3 ID and 6 OOD datasets) showed that our CLIPN outperforms existing methods. In summary, our contributions are Four-fold.

- We propose a novel CLIP architecture, named CLIPN, which equips CLIP with a “no” logic via the learnable “no” prompts and a “no” text encoder.
- We propose the image-text binary-opposite loss and text semantic-opposite loss, which teach CLIPN to match images with “no” prompts, thus learning to identify unknown samples.
- We propose two novel threshold-free inference algorithms (competing-to-win and agreeing-to-differ) to perform OOD detection via using negation semantics.
- Experimental results show that our CLIPN outperforms most existing OOD detection algorithms on both

large-scale and small-scale OOD detection tasks.

2. Related Work

2.1. Contrastive Vision-Language Models

One key research topic in artificial intelligence is studying the relationship between vision and language. Previously, many attention-based methods like BAN [17], Intra-Inter [11], and MCAN [47] dominate the visual-language tasks. Then, inspired by BERT [4], many methods [29, 39, 34] further contribute to the area via exploiting more advanced transformer architectures and prompt strategies. As transformer-based networks [6, 28] are successfully deployed into computer vision tasks, CLIP [35] is proposed to learn informative features from the vision-language pairs on top of the large-size model and datasets [36] via contrastive learning [5]. Further, [44, 46] demonstrate its impressive generalization to downstream tasks [23]. In addition, following the same lines as [34], there are also many works focusing on improving visual-language models via prompt engineering in computer vision. CoOp [50] and CoCoOp [49] work on changing the manual prompt to the learnable one for the purpose of better matching texts and images in a specific supervised task. Unlike these methods that provide more precise text descriptions or supervised guidance, this paper empowers the “no” logic within CLIP by adding learnable “no” prompts and text encoders during the unsupervised pre-trained period.

2.2. CLIP-based Zero-Shot Learning

Recently, zero-shot learning [15, 43] (ZSL) gains considerable interest. ZSL focuses on conferring the model with the ability to capture open (or unseen) knowledge. CLIP has contributed significantly to ZSL by virtue of CLIP’s excellent open-set versatility [46]. The representation space of CLIP is driven by a huge dataset [36] in an unsupervised manner and exhibits holistic, task-agnostic, and informative. These advantages [35] of CLIP are the primary factors that have led to its outstanding zero-shot performance. Further, the zero-shot adaptability of CLIP is found to work across many domains and different practical tasks [26, 44, 24, 22]. ZOC [8] employs an additional image-text dataset and bert [4] module to learn the names of OOD classes. MCM [30] capture the powerful ID classifier from class prompts and get better ID-ness from CLIP models. Unlike these methods, we exploit the open-world knowledge in CLIP to design a new OOD detection pipeline, where uses CLIP to straightly identify OOD samples via negation-semantic prompts.

2.3. Out-of-Distribution Detection

The goal of OOD detection is to detect OOD images from the test dataset (containing both ID and OOD images).

Designing the score function is the most popular method in OOD detection tasks. The scores are mainly derived from three sources: the probability, the logits, and the feature. For the probability, Hendrycks *et al.* [14] presented a baseline method using the maximum predicted softmax probability (MSP) as the ID score. Hendrycks *et al.* [13] also proposed to minimize KL-divergence between the softmax and the mean class-conditional distributions. For the logits, Hendrycks *et al.* [13] proposed using the maximum logit (MaxLogit) method. The energy score method [27] proposed to compute the logsumexp function over logits. For the feature, Lee *et al.* [21] computed the minimum Mahalanobis distance between the feature and the class centroids. Ndiour *et al.* [31] used the norm of the residual between the feature and the pre-image of its low-dimensional embedding. Furthermore, [1] proposes a solution that utilizes contrastive loss to distinguish between ID and OOD samples. In contrast to the methods mentioned earlier, which solely rely on the concept of ID-ness derived from images, our CLIPN can directly identify OOD samples by considering both images and negation-semantic texts.

3. Methodology

3.1. Preliminary: CLIP-based OOD Detection

Profiting from massive volumes of training data and large-size models, CLIP has demonstrated impressive performance in the zero-shot out-of-distribution classification task. Naturally, exploring the potential of CLIP on zero-shot out-of-distribution detection is worthwhile.

Referring to the existing CLIP-based zero-shot tasks [10, 30], we briefly review how to perform zero-shot OOD detection. The image encoder of CLIP is used to extract the image feature. Although there is no classifier in CLIP, we can use the names of ID classes to build the text inputs (e.g., “a photo of the dog”). Then obtain the text features from the text encoder as the class-wise weights that functionally play the same role as the classifier. Finally, taking the maximum softmax probability (MSP) algorithm as an example, we can calculate MSP according to the image feature and class-wise weights. The image will be treated as OOD if MSP is smaller than a pre-defined threshold, and vice versa.

3.2. Overview of CLIPN

In this paper, we customize a new CLIP architecture, named CLIP saying “no” (CLIPN), to exploit the potential of OOD detection in the original CLIP.

Architecture. As shown in Figure 3, our CLIPN is built on top of a pre-trained CLIP and consists of three modules: **(1)Image Encoder:** ϕ . The image encoder ϕ of CLIPN keeps the same structure and parameters as the image encoder of the pre-trained CLIP. In this paper, ViT-B [6] and ViT-L [6] are deployed to instantiate ϕ , respectively. By de-

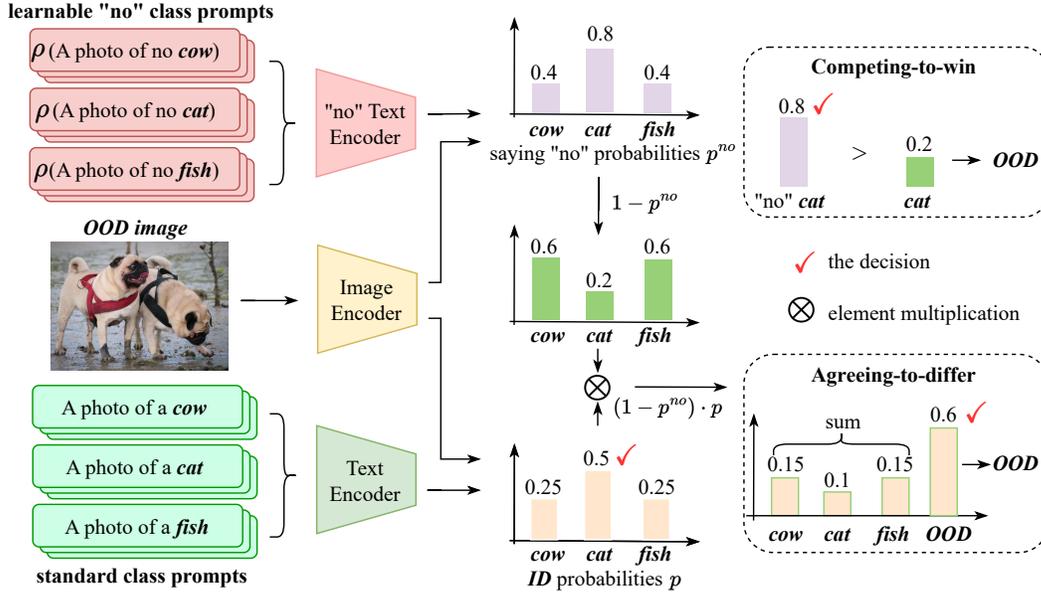


Figure 3. The inference pipeline of CLIPN. It consists of three networks: the image encoder, text encoder and “no” text encoder with learnable “no” prompts ρ . For the inference period, we propose competing-to-win and agreeing-to-differ to jointly determine the result with two text encoders. The ID classes are cow, cat, fish, and the OOD class is dog here.

fault, the parameters of ϕ are frozen. **(2) Text Encoders:** ψ . The text encoder ψ of CLIPN keeps the same structure and parameters as the text encoder of the pre-trained CLIP. The input of ψ also keeps the same, i.e., a standard text that describes one image (e.g., “a photo with/of/...”). **(3) “no” Text Encoders:** ψ^{no} . It is initialized by the text encoder of the pre-trained CLIP. But the difference from the ψ is that we set ψ^{no} learnable. The input of ψ^{no} is a negative text that describes one image with the opposite semantic.

Besides, we access the pre-trained CLIP from the OpenCLIP [16]. In this paper, we use two pre-trained models, including the CLIP based on ViT-B and ViT-L pre-trained using Laion-2B dataset [36], respectively.

Pre-training CLIPN. Obviously, the input texts of ψ and ψ^{no} are the opposite semantic in terms of one input image, just like the opposite attribute of ID and OOD. To teach CLIPN to distinguish which semantic is correctly matched with the image, we first design a novel “no” prompt strategy which introduces the textual descriptions with negative logic into ψ^{no} ; see Sec. 3.3 for details.

Based on the above-mentioned thought of opposite semantic, we propose an image-text binary-opposite loss and text semantic-opposite loss (See Sec. 3.4 for details) which allow CLIPN to learn explicitly when and how to align images with two kinds of texts.

Inference Stage. In the inference period, we further propose two novel threshold-free algorithms to determine if the input image is OOD (See Sec. 3.5 for details). The first is named the competing-to-win algorithm. It is summarized as follows: we first use the standard text and the input image to

predict the ID probabilities in terms of ID classes. Then employ the “no” text of the class with the highest probability to determine if the ID prediction is correct. The second is named the agreeing-to-differ algorithm. We use “no” texts to shrink the ID probabilities and generate a new probability of the OOD class. If the OOD probability is highest, the input image will be judged as OOD.

3.3. Prompt Design

We propose a new prompting strategy in the unsupervised pre-trained period: learnable “no” prompt pools. In the original CLIP, the image is depicted in positive logic, i.e., the content of the text is semantically consistent with the content of the image. We attempt to supplement a kind of negative logic, i.e., the content of the text is semantically negative with the content of the image. To this end, we define a series of “no” prompts to equip the original texts. Denoting the text for one image x is t , the “no” prompt pool is defined as $pool_{no}(t) = \{“a photo without \{t\}”, “a photo not appearing \{t\}”, \dots, “a photo not containing \{t\}”\}$, where has L handcrafted “no” prompts. The handcrafted “no” prompt pool is built upon [10] and modified with negative keywords. During training, given the input text t , we randomly pick a “no” prompt from $pool_{no}$ to generate the text with “no” logic, t^{no} . Next, a token embedding layer ρ is used to embed the “no” texts as a group of token feature vectors $\rho(t^{no})$, which are the input tokens of the “no” text encoder.

Moreover, inspired by [49], we design the learnable “no” prompts. We replace the negative keywords (e.g., “a photo

without”) by some learnable parameters σ that present the token features of negative semantics and are combined with the token features of t in the feature space.

3.4. Training Loss Design

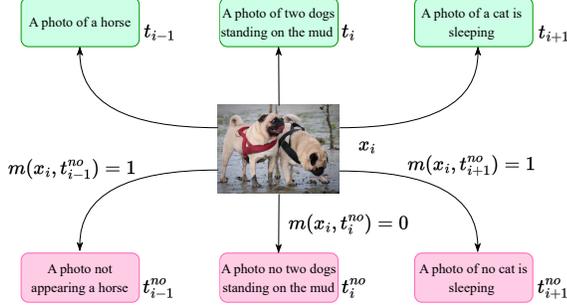


Figure 4. The illustration for matching x and t^{no} . The green and pink boxes indicate the standard text t and “no” text t^{no} , respectively. $m(x_i, t_j^{no}) = 1$ indicates they are *matched yet unrelated* (i.e., the “no” text is not a wrong description yet semantically irrelevant). $m(x_i, t_j^{no}) = 0$ indicates they are *reversed matched* (i.e., the “no” text has opposite semantic to the image).

In the training period, we define a mini-batch with N input pairs as $\mathcal{B} = \{(x_i, t_i, t_i^{no})\}_{i=1}^N \in \mathbb{D}_{clip}$ where (x_i, t_i, t_i^{no}) is the i -th input pair including an image, a standard text, and a “no” text. \mathbb{D}_{clip} is the dataset to pre-train CLIPN, such as CC12M [37]. Then the image feature f_i , text feature g_i and “no” text feature g_i^{no} can be calculated as follows:

$$\begin{aligned} f_i &= \phi(x_i) \\ g_i &= \psi(t_i) \\ g_i^{no} &= \psi^{no}(t_i^{no}), \end{aligned} \quad (1)$$

where $f_i, g_i, g_i^{no} \in \mathbb{R}^{1 \times D}$ and D is the feature dimension. All features are normalized by L_2 normalization operation.

We train CLIPN through two loss functions.

Image-Text Binary-Opposite Loss (ITBO). This loss function assists the model in matching the image feature with the correct “no” text features. To be specific, we define two relations between images and “no” texts: (1) matched yet unrelated (i.e., the “no” text is not the wrong description yet semantically irrelevant to the image); (2) reversed matched (i.e., the “no” text has opposite semantic to the image), as explained in Fig. 4. Consequently, the match-ness $m(x_i, t_j^{no})$ between the i -th image and the j -th “no” text can be defined as follow:

$$m(x_i, t_j^{no}) = m_{ij} = \begin{cases} 0, & i = j, \\ 1, & i \neq j, \end{cases} \quad (2)$$

where $m(x_i, t_j^{no}) = 0$ indicates they are reversed matched and $m(x_i, t_j^{no}) = 1$ indicates they are matched yet unrelated. Then we drive CLIPN to match images and “no”

texts in the feature space, guided by the match-ness. The loss \mathcal{L}_{itbo} is formulated as:

$$\begin{aligned} \mathcal{L}_{itbo}(\mathcal{B}) &= -\frac{1}{N} \sum_{i=1}^N (1 - m_{ii}) \log(1 - p_{ii}^{no}) \\ &\quad - \frac{1}{N(N-1)} \sum_{i=1}^N \sum_{j \neq i}^N m_{ij} \log p_{ij}^{no}, \end{aligned} \quad (3)$$

where p_{ij}^{no} presents the matched probability between the i -th image and j -th “no” text, formulated as follow:

$$p_{ij}^{no} = \frac{e^{\langle f_i, g_j^{no} \rangle / \tau}}{e^{\langle f_i, g_j \rangle / \tau} + e^{\langle f_i, g_j^{no} \rangle / \tau}}, \quad (4)$$

where \langle, \rangle indicates the inner product of two vectors and τ is a learnable temperature parameter.

Text Semantic-Opposite Loss (TSO). Obviously, the “no” prompts render t_i and t_i^{no} semantically opposite. Hence, in the feature space, the g_i and g_i^{no} also should be far from each other. To this end, the text semantic-opposite loss \mathcal{L}_{tso} is defined as:

$$\mathcal{L}_{tso}(\mathcal{B}) = \frac{1}{N} \left(2 - \sum_{i=1}^N \|g_i - g_i^{no}\|_2 \right), \quad (5)$$

where $\|\cdot\|_2$ is the L_2 distance function. When all g_i and g_i^{no} pairs are embedded into the opposite directions in the feature space, $\mathcal{L}_{tso}(\mathcal{B})$ will decrease to 0. The total loss in a mini-batch is calculated by adding the above two loss values: $\mathcal{L}(\mathcal{B}) = \mathcal{L}_{itbo}(\mathcal{B}) + \mathcal{L}_{tso}(\mathcal{B})$.

3.5. Inference algorithm of CLIPN

When deploying CLIPN to the test datasets of \mathbb{D}_{id} and \mathbb{D}_{ood} to perform a zero-shot OOD detection task, we only need all class names. For the C ID classes task, as shown in Figure 3, the name of each class is feed into standard and “no” texts, then fed into ψ, ψ^{no} . Actually, we introduce the prompt pools in Sec. 3.3. A prompt ensemble strategy is employed to generate the text features via adding the text features of all prompts together, followed [10] Then the ID class probability can be formulated as:

$$p_{ij} = \frac{e^{\langle f_i, g_j \rangle / \tau}}{\sum_{k=1}^C e^{\langle f_i, g_k \rangle / \tau}}, \quad (6)$$

where C is the number of classes and p_{ij} presents the predicted probability for the input image x_i belonging to the j -th class. The matched probability p_{ij}^{no} between the image and the j -th “no” class text can be calculated via Eqn. 4. Next, we propose two novel threshold-free algorithms to determine if x_i is OOD.

Competing-to-win Algorithm. The first algorithm is named the competing-to-win (CTW). Motivated by MSP,

we find the class with the maximum ID probability. Then we compare the value of p^{no} and p^{yes} (i.e., $1 - p^{no}$) to determine if x_i is OOD (or ID). The above process can be formulated as follows:

$$I_{ctw} = 1 - p_{ij}^{no}, j = \arg \max\{p_{ik}\}_{k=1\sim C}$$

$$\mathbb{I}(x_i) = \begin{cases} 1, & 1 - p_{ij}^{no} \geq p_{ij}^{no}, \\ 0, & \text{else,} \end{cases} \quad (7)$$

where j indicates the class with the highest ID probability, I_{ctw} is the ID-ness, and $\mathbb{I}(x_i)$ is an indicator that presents x_i is an ID image when it is equal to 1, while is OOD.

Agreeing-to-differ Algorithm. Nevertheless, the above strategy is slightly aggressive. It will fail in some hard-distinguish situations (e.g., the maximum ID probability is not significantly higher than other probabilities). To make the decision flexible, we propose another algorithm, named agreeing-to-differ (ATD), to take all ID probabilities and p^{no} into account. It can re-formulate the C -classes probabilities as the $(C + 1)$ -classes probabilities. An unknown class will be created, and its probability is defined as:

$$p_{C+1} = 1 - \sum_{j=1}^C (1 - p_{ij}^{no}) p_{ij}. \quad (8)$$

The OOD sample can be detected as:

$$I_{atd} = 1 - p_{C+1}$$

$$\mathbb{I}(x_i) = \begin{cases} 1, & p_{C+1} \leq \max\{p_{ij}\}_{j=1\sim C}. \\ 0, & \text{else.} \end{cases} \quad (9)$$

If the probability of the unknown class p_{C+1} is larger than all ID probabilities, the input image will be detected as OOD. Otherwise, it is determined as ID.

4. Experiment

4.1. Experimental Details

ID and OOD Datasets. In this section, we evaluate the performance of our approach and compare it to state-of-the-art OOD detection algorithms. We focus on three different OOD detection tasks.

(1) OOD detection on large-scale datasets. Following the prior work [30] on large-scale OOD detection, we choose ImageNet-1K [20] as the ID dataset. Four OOD datasets (including Texture [42], iNaturalist [41], SUN [45], and Places365 [48]) are used to comprehensively benchmark the algorithms.

(2) OOD detection on small-scale datasets. In this setting [18], CIFAR-100 [19] is set as an ID dataset, and the OOD datasets are CIFAR-10 [19], ImageNet_R [18], and LSUN [18]. The data scale of CIFAR-100 is significantly smaller than ImageNet-1K.

(3) OOD detection on the in-domain dataset. In addition, ZOC [8] is the first work on the zero-shot OOD detection task. To compare with it, we follow its experimental setting and test our method on CIFAR-10 and CIFAR-100 datasets. The reported performance is averaged over 5 splits.

Evaluation Metrics. We report two widely used metrics, AUROC and FPR95, for performance evaluation. AUROC is a metric that computes the area under the receiver operating characteristic curve. A higher value indicates better detection performance. FPR95 is short for FPR@TPR95, which is the false positive rate when the true positive rate is 95%. Smaller FPR95 implies better performance.

Model Details. All used pre-trained CLIP models are obtained from OpenCLIP [16]. The CLIP based on ViT-B-16 and ViT-B-32 are used for experiments.

Previous Methods for Comparison. We compare our method with 5 prior logit-based works on OOD detection and 2 works on zero-shot OOD detection. They are MSP [14], Energy [27], MaxLogit [13], ReAct [38], ODIN [25], ZOC [8] (only on the third task) and MCM [30]. For evaluation, ID-ness scores proposed by the above works and this paper (see Eqn. 7 and 9) are used for the calculation of AUROC and FPR95.

Reproduction details. CLIPN is pre-trained on the CC-3M [37] dataset for ten epochs, using a batch size of 2048, and trained on a system equipped with 4×NVIDIA RTX3090 (24G) and 2×Intel Xeon Gold 5118. All experiments involving our proposed algorithms are independently repeated three times, and the reported results are averaged from these three experiments. To ensure fairness and versatility in our evaluation, the model used for performance assessment is the final-epoch model, rather than selecting the optimal model from all epochs.

4.2. Results on Zero-shot OOD Detection

OOD detection on large-scale datasets. Large-scale OOD detection contributes to real-world applications. We give a comprehensive OOD evaluation on ImageNet-1K in Table 1. Our proposed method, CLIPN-A with learnable 'no' texts, outperforms the previous state-of-the-art method MCM [30] across all OOD datasets, achieving the highest AUROC and FPR95 scores. On average, our approach using ViT-B-16 demonstrates significant enhancements of at least 2.34% and 11.64% in terms of AUROC and FPR95, respectively. Similarly, our approach utilizing ViT-B-32 showcases improvements of at least 4.12% and 23.25% on AUROC and FPR95, respectively, underscoring its superior performance.

OOD detection on small-scale datasets. In Table 2, we present the results of our method on CIFAR-100. Our approach achieves the best average performance on small datasets in terms of AUROC and FPR95. Overall, our experiment yielded an impressive conclusion: our algorithms

Table 1. Results on large-scale zero-shot OOD detection. ID dataset is ImageNet-1K. All experimental results are evaluated using CLIP model based on ViT-B-16 and ViT-B-32. CLIPN-C and CLIPN-A indicate CLIPN with competing-to-win and agreeing-to-differ algorithms, respectively. The number with **red** and **green** indicates the improved AUROC and FPR95 of our algorithms compared to the baseline algorithms, respectively. The **bold** number presents the best performance on each dataset in terms of AUROC and FPR95. The result of MCM is reported on [30] and the results of other compared methods are reproduced by us.

Method	iNaturalist		SUN		Texture		Places		Avg	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
Image Encoder: ViT-B-16										
MSP [14]	77.74	74.57	73.97	76.95	74.84	73.66	72.18	79.72	74.68	76.22
MaxLogit [13]	88.03	60.88	91.16	44.83	88.63	48.72	87.45	55.54	88.82	52.49
Energy [27]	87.18	64.98	91.17	46.42	88.22	50.39	87.33	57.40	88.48	54.80
ReAct[38]	86.87	65.57	91.04	46.17	88.13	49.88	87.42	56.85	88.37	54.62
ODIN [25]	57.73	98.93	78.42	88.72	71.49	85.47	76.88	87.8	71.13	90.23
MCM [30]	94.61	30.91	92.57	37.59	86.11	57.77	89.77	44.69	90.76	42.74
CLIPN-C (Ours)	90.88	28.58	89.38	31.64	78.28	56.59	86.85	37.55	86.35	38.59
CLIPN-A (Ours)	95.27	23.94	93.93	26.17	90.93	40.83	92.28	33.45	93.10 (+2.34)	31.10 (-11.64)
Image Encoder: ViT-B-32										
MSP [14]	77.14	74.62	71.86	80.64	73.76	74.88	70.61	82.09	73.34	78.06
MaxLogit [13]	84.01	75.05	88.27	56.59	84.41	60.00	85.68	61.65	85.59	63.32
Energy [27]	82.51	80.36	88.14	59.96	83.66	63.18	85.44	64.61	84.94	67.03
ReAct[38]	84.00	76.94	88.05	59.27	88.16	53.31	85.41	64.42	86.41	63.49
ODIN [25]	44.57	99.60	76.63	94.39	71.85	89.88	75.84	91.90	67.22	93.94
CLIPN-C (Ours)	87.24	38.42	84.49	43.43	72.93	65.41	76.26	60.73	80.23	52.00
CLIPN-A (Ours)	94.67	28.75	92.85	31.87	86.93	50.17	87.68	49.49	90.53 (+4.12)	40.07 (-23.25)

Table 2. Results on small-scale zero-shot OOD detection. ID dataset is CIFAR-100. The image encoder is ViT-B-32.

	CIFAR-10		ImageNet_R		LSUN		Avg	
	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
MSP [14]	77.83	80.02	85.09	58.42	82.51	68.88	81.81	69.11
MaxLogit [13]	85.20	57.58	78.79	83.72	89.02	61.21	84.34	67.51
Energy [27]	84.04	59.16	74.65	88.81	87.37	68.87	82.02	72.28
ReAct[38]	83.32	61.13	75.24	87.88	88.59	63.12	82.38	70.71
ODIN [25]	69.38	91.57	87.30	48.65	93.32	32.21	83.33	57.48
CLIPN-C (Ours)	85.44	49.84	82.72	65.79	91.51	44.26	86.56	53.30
CLIPN-A (Ours)	88.06	47.99	87.09	60.07	93.55	35.19	89.57 (+5.23)	47.75 (-9.73)

Table 3. Results of OOD detection on the in-domain setting. † and ‡ indicate that the results are evaluated based on pre-trained model in ZOC [8] and pre-trained model in our paper. Δ indicates the improved AUC and FPR95.

	CIFAR10		CIFAR100	
	AUROC	FPR95	AUROC	FPR95
CLIP+MSP † [14]	88.0±3.3	-	78.1±3.1	-
ZOC † [8]	93±1.7	-	82.1±2.1	-
Δ	+5	-	+4	-
CLIP+MSP ‡ [14]	85.58±2.83	68.68±8.86	74.76±4.42	84.29±6.78
CLIPN-C ‡	89.15±1.16	53.43±2.07	81.29±2.51	66.91±5.58
CLIPN-A ‡	90.08±1.14	53.21±2.53	82.51±1.78	65.98±5.34
Δ	+4.5	-15.47	+7.75	-18.31

demonstrate superior generalizability, consistently achieving optimal average performance for zero-shot OOD detection tasks of any scale.

OOD detection on in-domain datasets. Table. 3 shows

our results in the in-domain setting, where CIFAR-10 and CIFAR-100 are divided into ID groups and OOD groups, respectively. It’s worth noting that our method outperforms MSP by 4.5% and 7.75% AUROC on CIFAR10 and CIFAR100, respectively. Meanwhile, ZOC also outperforms MSP by 5% and 4% AUROC. However, it should be pointed out that our implemented MSP yielded lower scores than those reported in ZOC [8], which is caused by the higher-performance pre-trained model (but no public available pre-trained model of ZOC).

4.3. Ablation Study

The effectiveness of two losses. We conduct experiments to demonstrate the effectiveness of the proposed two loss functions, image-text binary-opposite (ITBO) loss and text semantic-opposite (TSO) loss. We train CLIPN using ITBO and ITBO+TSO, respectively. The results evaluated by

Table 4. Ablation study on the proposed loss. The reported AUC is evaluated by CLIPN-A with learnable “no” texts. ITBO and TSO indicate image-text binary-opposite loss and text semantic-opposite loss.

Loss	iNaturalist	SUN	Texture	Places
ITBO	93.08	91.61	89.52	89.83
ITBO + TSO	95.27	93.93	90.93	92.28

Table 5. Ablation study on the handcrafted and learnable “no” texts. \star and \dagger refer to our method uses the handcrafted and learnable “no” prompts, respectively. CLIPN-C and CLIPN-A indicate CLIPN with competing-to-win and agreeing-to-differ algorithms, respectively.

Model	iNaturalist	SUN	Texture	Places
CLIPN-C \star	84.05	82.53	77.62	79.99
CLIPN-C \dagger	90.88	89.38	78.28	86.85
Δ	6.83	6.85	0.66	6.86
CLIPN-A \star	92.75	91.55	92.15	89.65
CLIPN-A \dagger	95.27	93.93	90.93	92.28
Δ	2.52	2.38	-1.22	2.63

CLIPN-A with learnable “no” texts are reported in Table. 4. When solely utilizing ITBO, CLIPN-A achieves AUROC scores of 93.08%, 91.61%, 89.52%, and 89.83% across the four OOD datasets. Upon incorporating both ITBO and TSO, an additional enhancement of 2.19%, 2.32%, 1.41%, and 2.45% is observed in the AUROC scores.

The effectiveness of the handcrafted and learnable “no” texts. We conducted experiments to demonstrate the effectiveness of both handcrafted and learnable “no” texts. The handcrafted “no” texts are constructed by modifying the text prompts of CLIP using negative keywords. The number of learnable “no” keywords is set to 16. The results evaluated using CLIPN-C and CLIPN-A are presented in Table. 5. For CLIPN-C, the utilization of learnable “no” texts leads to improvements of AUROC by 6.83%, 6.85%, 0.66%, and 6.86% across the four datasets, compared to handcrafted texts. Regarding CLIPN-A, the incorporation of learnable “no” texts results in AUROC improvements of 2.52%, 2.38%, and 2.63% for three datasets, while showing a decrease of 1.22% AUROC for the Texture dataset, as compared to handcrafted texts.

The effectiveness to eliminate the mis-classified OOD samples. We evaluate MSP and CLIPN-A on CIFAR-10 (ID dataset) and CIFAR-100 (OOD dataset), then use T-SNE [40] to visualize the features (using Eqn. 1) of ID samples and mis-classified OOD samples (belonging to OOD yet classified as ID) caused by MSP and ours. The results are shown in Figure. 5. For MSP, the numbers of mis-classified OOD samples are 666, 410, and 171 under the threshold 0.5, 0.7, and 0.9, which are heavily falling into

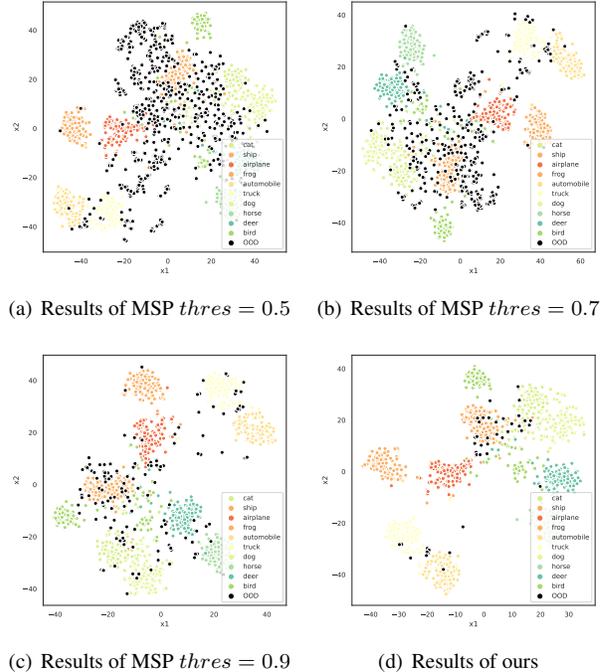
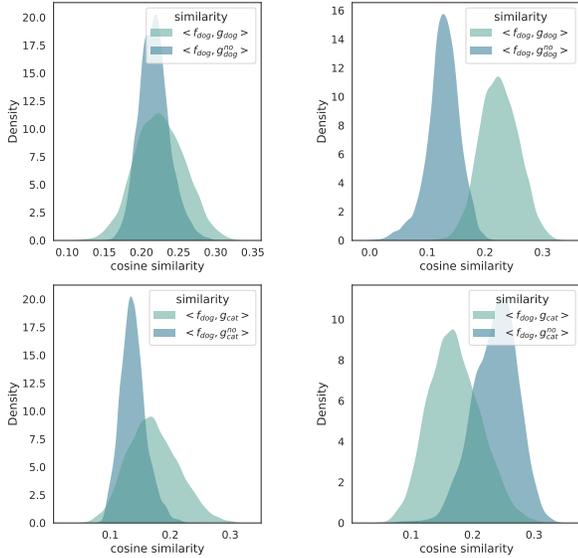


Figure 5. Mis-classified OOD (dark dots) and ID (other colors) feature visualization via T-SNE. (a-c) Results of MSP [14] with threshold [0.5, 0.7, 0.9]. (d) Result of our CLIPN-A. The ID and OOD features are captured from images in CIFAR-10 and CIFAR-100, using Eqn. 1. The misclassified samples indicate that samples were classified incorrectly as CIFAR-10 classes, either by MSP or by ours. Besides, the number of mis-classified OOD features are 666, 410, 171, and 45 for (a), (b), (c), and (d), respectively.

ID sample clusters. However, our threshold-free CLIPN-A only has 45 mis-classified OOD samples, significantly less than MSP and rarely fall into ID clusters. It demonstrates that our CLIPN-A can effectively use the “no” logic to eliminate mis-classified OOD samples.

The capacity to match images and “no” texts. We conduct an experiment to evaluate the capacity of matching images and “no” texts in terms of CLIP and CLIPN. Specifically, given CLIP or CLIPN, we extract all dog features from CIFAR-10 (for each dog image, we get the image feature f_{dog} using Eqn. 1). Then we use Eqn. 1 to get four text features, g_{dog} , g_{cat} (standard dog/cat texts) and g_{dog}^{no} , g_{cat}^{no} (“no” dog/cat texts). Next we calculate the two group cosine similarities $\langle f_{dog}, g_{dog} \rangle$ and $\langle f_{dog}, g_{dog}^{no} \rangle$; $\langle f_{dog}, g_{cat} \rangle$ and $\langle f_{dog}, g_{cat}^{no} \rangle$. Finally, we use the kernel density estimation function [33] to estimate the density of two group similarities for CLIP and CLIPN, as shown in Figure. 6. Followed the match-ness defined in Eqn. 2, the ideal outcome should be that dog features are matched with dog texts and “no” cat texts. From Figure. 6, we find that the original CLIP fails to achieve it because all similarities are mixed together. Conversely, similarity density of



(a) Similarity density of CLIP (b) Similarity density of CLIPN

Figure 6. Similarity density of CLIP and CLIPN. (a) and (b) show the similarity density between dog images in CIFAR-10 and four kind of texts in terms of CLIP and CLIPN, respectively.

CLIPN has obviously higher value on $\langle f_{dog}, g_{dog} \rangle$ and lower $\langle f_{dog}, g_{cat}^{no} \rangle$, implying CLIPN has strong ability to match images and “no” texts.

Computational and storage cost of the training and inference periods. Our CLIPN significantly reduces both training time and GPU memory usage when compared to training CLIP, all while achieving substantial performance improvements with minimal additional cost. Specifically, we employ four commonly used metrics to quantify these benefits: (1) Floating-Point Operations per Second (FLOPs); (2) Number of Parameters; (3) Training Time per Iteration; (4) GPU Memory Usage per Iteration. Metric (1) is calculated solely during the forward step, while metrics (3) and (4) pertain to the computation and memory usage during both the forward and backward steps.

In terms of the training period, as indicated in Table 6, CLIPN demonstrates reductions of 36.4% and 29.0% in training time per iteration and GPU usage on ViT-B-32, respectively. Additionally, on ViT-L-14, CLIPN achieves reductions of 58.6% and 65.0% in training time per iteration and GPU usage when compared to CLIP. This outcome is attributed to the decreased number of Floating-Point Operations (FLOPs) and parameters during backward computation. Furthermore, CLIPN enables the training of ViT-L-14 on the RTX3090 with a comparatively large batch size. Regarding the inference period, outlined in Table 7, the conversion of two text encoders to classifiers results in a negligible increase in computation and parameters, specifically by 0.016G and 0.51M, respectively. However, this minor

Table 6. Cost comparison between CLIP and CLIPN during the training period. Red presents the FLOPs and parameters that need backward computations. FLOPs/parameters are decomposed into sub-parts of image, text, and “no” text encoders. Blue indicates the percentage of the decreased cost. All experiments are evaluated on a device with 4×RTX3090 (24G) and 2×Intel Xeon Gold 5118.

Method	FLOPs (G)	Parameters (M)	Time/Iter (s)	GPU Usage (G)
ViT-B-32 (Input size: 16x3x224x224)				
CLIP	141 + 95	87.9 + 37.8	0.22	5.86
CLIPN (ours)	141 + 95 + 95	87.9 + 37.8 + 37.8	0.14 (↓ 36.4%)	4.16 (↓ 29.0%)
ViT-L-14 (Input size: 16x3x224x224)				
CLIP	2594 + 213	304.0 + 85.1	0.58	18.91
CLIPN (ours)	2594 + 213 + 213	304.0 + 85.1 + 85.1	0.24 (↓ 58.6%)	6.62 (↓ 65.0%)
ViT-L-14 (Input size: 64x3x224x224)				
CLIP	10375 + 852	304.0 + 85.1	-	OOM (> 24)
CLIPN (ours)	10375 + 852 + 852	304.0 + 85.1 + 85.1	0.49	8.31

Table 7. Cost and performance comparison between CLIP and CLIPN during the inference period. The performance of AUROC is CLIP with MaxLogits and CLIPN-A. The costs are calculated under the same set of ViT-B-32 as Table. 6.

Method	FLOPs (G)	Parameters (M)	Avg AUROC	Improved AUROC
CLIP	141 + 0.016	87.9 + 0.51	85.59%	-
CLIPN (ours)	141 + 0.016 + 0.016	87.9 + 0.51 + 0.51	90.53%	↑ 4.94%

adjustment leads to an average AUROC increase of 4.94%.

5. Conclusion and Limitation

This paper presents a novel framework, namely CLIPN, for OOD detection by teaching CLIP to say “no”. The key insight is to equip CLIP with the capability of distinguishing OOD and ID samples via positive-semantic prompts and negation-semantic prompts. Specifically, we propose new “no” prompts and text encoder. Further, we propose two training losses: an image-text binary-opposite loss and a text semantic-opposite loss. These losses enable CLIP to recognize scenarios where it should respond with “no” and to understand the meaning of “no”. Additionally, we propose two threshold-free inference algorithms: competing-to-win and agreeing-to-differ. Extensive experimental results demonstrated the effectiveness of our method. One limitation of our approach is the lack of clear demonstrations of its extension to OOD segmentation or detection tasks. Another limitation of our approach is the uncertainty regarding its effectiveness for OOD classification in specialized datasets such as medical and satellite images. This is primarily because our model is based on CLIP, and its effectiveness in specialized datasets is still underexplored.

6. Acknowledgement

This work is supported by grants from Foshan HKUST Projects (Grants FSUST21-HKUST10E and FSUST21-HKUST11E), the Hong Kong Innovation and Technology Fund (Projects ITS/030/21), and the Beijing Institute of Collaborative Innovation (BICI) in collaboration with HKUST (Grant HCIC-004).

References

- [1] Jianhong Bai, Zuo Zhu Liu, Hualiang Wang, Jin Hao, Yang Feng, Huanpeng Chu, and Haoji Hu. On the effectiveness of out-of-distribution data in self-supervised long-tail learning. *arXiv preprint arXiv:2306.04934*, 2023. 3
- [2] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016. 1
- [3] Jinghui Chen, Saket Sathe, Charu Aggarwal, and Deepak Turaga. Outlier detection with autoencoder ensembles. In *Proceedings of the 2017 SIAM international conference on data mining*, pages 90–98. SIAM, 2017. 1
- [4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3
- [5] Xinpeng Ding, Nannan Wang, Shiwei Zhang, De Cheng, Xiaomeng Li, Ziyuan Huang, Mingqian Tang, and Xinbo Gao. Support-set based cross-supervision for video grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11573–11582, 2021. 3
- [6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 3
- [7] Nick Drummond and Rob Shearer. The open world assumption. In *eSI Workshop: The Closed World of Databases meets the Open World of the Semantic Web*, volume 15, page 1, 2006. 1
- [8] Sepideh Esmailpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pretrained model clip. In *Proceedings of the AAAI conference on artificial intelligence*, 2022. 2, 3, 6, 7
- [9] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081, 2021. 1
- [10] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 2, 3, 4, 5
- [11] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6639–6648, 2019. 3
- [12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1
- [13] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019. 1, 2, 3, 6, 7
- [14] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 1, 2, 3, 6, 7, 8
- [15] Dat Huynh and Ehsan Elhamifar. Fine-grained generalized zero-shot learning via dense attribute-based attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4483–4493, 2020. 3
- [16] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hananeh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below. 4, 6
- [17] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. *Advances in neural information processing systems*, 31, 2018. 3
- [18] Rajat Koner, Poulami Sinhamahapatra, Karsten Roscher, Stephan Günnemann, and Volker Tresp. Oodformer: Out-of-distribution detection transformer. *arXiv preprint arXiv:2107.08976*, 2021. 6
- [19] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6
- [20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 2, 6
- [21] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 3
- [22] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023. 3
- [23] Yi Li, Hualiang Wang, Yiqun Duan, Hang Xu, and Xiaomeng Li. Exploring visual interpretability for contrastive language-image pre-training. *arXiv preprint arXiv:2209.07046*, 2022. 3
- [24] Yi Li, Huifeng Yao, Hualiang Wang, and Xiaomeng Li. Freeseq: Free mask from interpretable contrastive language-image pretraining for semantic segmentation. *arXiv preprint arXiv:2209.13558*, 2022. 3
- [25] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017. 1, 6, 7
- [26] Jiayang Liu, Tianxiang Hu, Yan Zhang, Xiaotang Gai, Yang Feng, and Zuo Zhu Liu. A chatgpt aided explainable framework for zero-shot medical image diagnosis. *arXiv preprint arXiv:2307.01981*, 2023. 3
- [27] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020. 1, 2, 3, 6, 7

- [28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 3
- [29] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3
- [30] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyu Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *arXiv preprint arXiv:2211.13445*, 2022. 2, 3, 6, 7
- [31] Ibrahima Ndiour, Nilesh Ahuja, and Omesh Tickoo. Out-of-distribution detection with subspace techniques and probabilistic modeling of features. *arXiv preprint arXiv:2012.04250*, 2020. 3
- [32] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015. 1
- [33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 8
- [34] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019. 3
- [35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3
- [36] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 2, 3, 4
- [37] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018. 5, 6
- [38] Yiyu Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021. 6, 7
- [39] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 3
- [40] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8
- [41] Grant Van Horn, Oisín Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 6
- [42] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4921–4930, 2022. 1, 6
- [43] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6857–6866, 2018. 3
- [44] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022. 3
- [45] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 6
- [46] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 3
- [47] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290, 2019. 3
- [48] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 6
- [49] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 3, 4
- [50] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3