

# Equipping Vision Foundation Model with Mixture of Experts for Out-of-Distribution Detection

Shizhen Zhao Jiahui Liu Xin Wen Haoru Tan Xiaojuan Qi<sup>†</sup>  
The University of Hong Kong

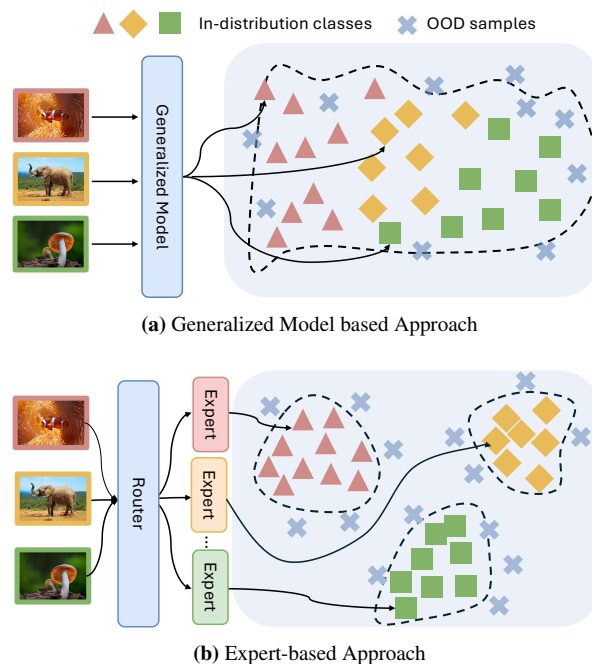
## Abstract

Pre-trained vision foundation models have transformed many computer vision tasks. Despite their strong ability to learn discriminative and generalizable features crucial for out-of-distribution (OOD) detection, their impact on this task remains underexplored. Motivated by this gap, we systematically investigate representative vision foundation models for OOD detection. Our findings reveal that a pre-trained DINOv2 model, even without fine-tuning on in-domain (ID) data, naturally provides a highly discriminative feature space for OOD detection, achieving performance comparable to existing state-of-the-art methods without requiring complex designs. Beyond this, we explore how fine-tuning foundation models on in-domain (ID) data can enhance OOD detection. However, we observe that the performance of vision foundation models remains unsatisfactory in scenarios with a large semantic space. This is due to the increased complexity of decision boundaries as the number of categories grows, which complicates the optimization process. To mitigate this, we propose the Mixture of Feature Experts (MoFE) module, which partitions features into subspaces, effectively capturing complex data distributions and refining decision boundaries. Further, we introduce a Dynamic- $\beta$  Mixup strategy, which samples interpolation weights from a dynamic beta distribution. This adapts to varying levels of learning difficulty across categories, improving feature learning for more challenging categories. Extensive experiments demonstrate the effectiveness of our approach, significantly outperforming baseline methods. The project will be available at [shizhen-zhao.github.io/OOD\\_MoFE/](https://shizhen-zhao.github.io/OOD_MoFE/).

## 1. Introduction

The task of out-of-distribution (OOD) detection [16, 26, 31, 49] aims to equip models with the capability to discern whether input images originate from unknown OOD classes or belong to in-domain (ID) classes. Mainstream OOD detection methods [9, 10, 25, 63] focus on learning features

<sup>†</sup>Corresponding author



**Figure 1. Holistic comparison to previous philosophy.** (a) Traditional methods use a generalized model to project inputs onto a complex distribution; (b): Our approach leverages multiple experts to break the complex distribution into smaller ones, which leads to compact ID distribution and simplified decision boundary.

and classifiers [31, 52, 53, 68] from ID data and then develop a score metric [14, 15, 35, 54] to determine whether a sample belongs to ID or OOD classes. Despite significant advancements, the fundamental challenge in OOD detection is establishing a feature space with high discriminative capacity that can effectively distinguish OOD samples from ID samples. Recently, vision foundation models [21, 38, 46, 51] trained on large-scale datasets have demonstrated the ability to learn robust and generalizable features, benefiting numerous tasks [29, 65, 76, 78]. This raises the question: *with such powerful models and feature representations, does OOD detection remain a problem?*

Although several studies [11, 40, 41, 70] have explored

the use of foundation models for OOD detection, most focus on improving the performance of vision-language models like CLIP [46], while other foundation models, such as DINOv2, remain largely unexamined. In this study, we systematically investigate the feature spaces of different representative pre-trained foundation models, including vision-language models (e.g., CLIP) and self-supervised models (e.g., DINOv2), in the context of OOD detection. Our results reveal that DINOv2 provides the most discriminative feature space, enabling effective OOD detection without any fine-tuning. Notably, using a simple KNN metric [54], DINOv2 achieves performance comparable to more complex methods, establishing a strong baseline for further research.

While vision foundation models [38, 45, 46, 70] have achieved impressive performance in OOD detection, there is still room for improvement, particularly on in-domain data with large semantic spaces [57–61, 71] (e.g., 29.27% FPR95 on the ImageNet-1K OOD benchmark [54]). This prompts us to investigate whether foundation models can be further optimized by leveraging available ID data. However, as the number of semantic classes increases, the complexity of the decision boundaries required to distinguish between ID and OOD data grows as well [17, 28, 32]. This heightened complexity creates challenges when fine-tuning foundation models on limited ID data. Previous methods (e.g. MOS [17]) decouple the complex space into simpler subspaces from the perspective of loss, which eases the optimization process and simplifies the decision boundaries. In this study, we tackle the problem orthogonally from the model perspective by designing a new Mixture of Experts (MoE) architecture to more thoroughly disentangle complex ID distribution.

To address this issue, rather than directly optimizing the whole feature space [3, 33, 34, 37] (Fig. 1(a)), we propose a Mixture-of-Feature-Expert (MoFE) module, which utilizes multiple experts, and each expert specializes in a specific subspace and optimizes it accordingly (Fig. 1(b)). MoFE operates by partitioning the original feature space into  $K$  subspaces based on semantics and feature similarities within the ID dataset. Each subspace is assigned to a dedicated expert, and a router assigns samples to the appropriate expert based on these partitions. Different from previous studies [12, 50], we use the [CLS] token as the input to the router network, since it encapsulates the semantic feature of the whole image. During training, the router is supervised by the partition assignments to ensure accurate sample-to-expert mapping. Each expert focuses solely on optimizing features within its designated partition, which helps prevent interference between features from different partitions. The results show that our approach significantly surpasses the previous approaches by a large margin (see Tab. 1), revealing the importance of learning expert models for OOD tasks.

Additionally, given that data augmentation has been shown to enhance generalization for OOD, we introduce

a novel Mixup data augmentation strategy to further improve feature learning, which is better suited for advanced vision foundation models. Our design is based on the observation that different categories exhibit varying levels of discriminativeness with features from vision foundation models. In the original feature space, some categories show high discriminativeness, while others do not. For categories that are already well-represented, synthesizing dissimilar samples via vanilla Mixup [64] can blur the decision boundary between ID and OOD, leading to degraded performance (Appendix Fig. 6). Thus, unlike existing Mixup strategies that treat all categories equally [56, 62, 72], our approach makes Mixup weight sampling category-dependent by adjusting the sampling distribution (*i.e.* beta distribution) dynamically, taking into account their discriminativeness.

Our major **contributions** can be summarized as follows:

- We design a novel MoFE module to tailor pre-trained vision foundation models for OOD detection. This approach reduces the difficulty of fitting complex data distributions from limited data and eases the optimization process.
- We explore the effectiveness of the raw feature spaces from various vision foundation models for OOD detection. Through analysis, we leverage DINOv2 with simple scoring metric to establish a strong baseline. Additionally, we designed a Dynamic- $\beta$  Mixup that is better suited for advanced vision foundation models.
- Our extensive experimental results demonstrate the effectiveness of the proposed model, achieving significant improvements over several competitive baseline methods on standard benchmarks.

## 2. Related Work

**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 [14, 15], the logits [15, 35], and the feature [26, 44]. Some studies [20, 49, 69] focus on leveraging contrastive learning to enhance the feature representation. Other studies show that synthesizing pseudo samples [9, 49, 55, 55] as OOD instances is also a promising approach to make the feature space more compact. The methods [17, 28, 32] are the most relevant to ours, which also break the semantic space into smaller ones. Different from these approaches, in our design, we propose a novel MoE module, with each expert exclusively concentrating on optimizing features within its specific partition. Our results demonstrate that our approach outperforms them by a substantial margin.

**OOD Detection with Foundation Models** There are some existing OOD detection methods [11, 40–42, 45, 70] leveraging foundation models. Maximum Concept Matching

(MCM) [41] proposes a simple yet effective zero-shot OOD detection method by aligning visual features with textual concepts. Some other studies [45, 70] explore negative prompts to learn the diversity of negative features, enabling more accurate detection of OOD samples. Although these studies have made great progress by leveraging CLIP to enhance the performance in existing benchmarks, they only explore and fine-tune CLIP. In our studies, we explore different foundation models and explore a better fine-tuning paradigm.

**Mixture of Experts** Mixture of Experts has been studied independently in both computer vision [36, 43, 47, 73] and natural language processing [12, 22, 27, 50]. These works are studied in the context of conditional computation, which is to increase the number of model parameters without a proportional increase in computational cost. Currently, some studies [5, 24] explore improving expert specialization and leveraging MoE to mitigate data conflict problems, where some data might interfere with each other. In our study, we introduce MoFE to the out-of-distribution task in the context of foundation models and build specialized OOD detectors for different feature subspaces.

### 3. Pilot Study

In this section, we first introduce preliminaries for the OOD detection task in Sec. 3.1. Then, we explore the impact of foundation models on OOD detection performance and analyze their strengths and weaknesses in Sec. 3.2.

#### 3.1. Preliminaries

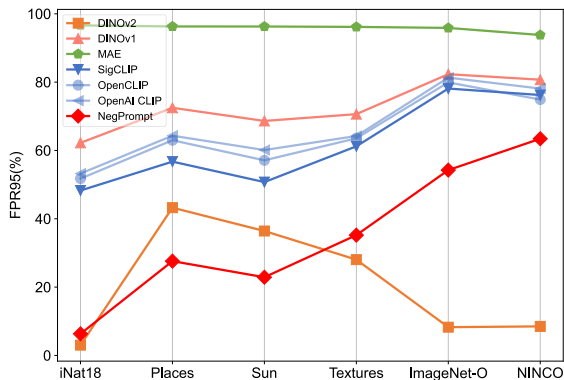
We consider supervised multi-class classification, where  $\mathcal{X}$  represents the input image space and  $\mathcal{Y} = \{1, 2, \dots, C\}$  represents the label space. The training dataset  $\mathbb{D}_{in} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$  is drawn independently and identically distributed (*i.i.d.*) from the joint data distribution  $P_{\mathcal{X}\mathcal{Y}}$ . Let  $\mathcal{P}_{in}$  denote the marginal distribution on  $\mathcal{X}$ . Let  $f : \mathcal{X} \mapsto \mathbb{R}^{|\mathcal{Y}|}$  be a neural network trained on samples drawn from  $P_{\mathcal{X}\mathcal{Y}}$  to output a logit vector, which is used to predict the label of the input sample.

**Out-of-distribution Detection.** When deploying a machine learning model in real-world scenarios, it is crucial for a reliable classifier not only to accurately classify known in-distribution (ID) samples, but also to recognize any out-of-distribution (OOD) inputs as "unknown". This can be accomplished by incorporating an OOD detector alongside the classification model  $f$ . OOD can be formulated as a binary classification task. During testing, the objective is to determine whether a sample  $\mathbf{x} \in \mathcal{X}$  belongs to  $\mathcal{P}_{in}$  (ID) or not (OOD). This decision can be made using a scoring

metric  $S(\mathbf{x})$ :

$$G_\lambda(x) = \begin{cases} \text{ID} & S(\mathbf{x}) \geq \lambda \\ \text{OOD} & S(\mathbf{x}) < \lambda \end{cases}, \quad (1)$$

where samples with higher scores  $S(\mathbf{x})$  are classified as ID and vice versa, and  $\lambda$  is the threshold. Some typically used metrics  $S(\mathbf{x})$  include MSP [14], MaxLogit [15], Energy [35] and KNN [54].

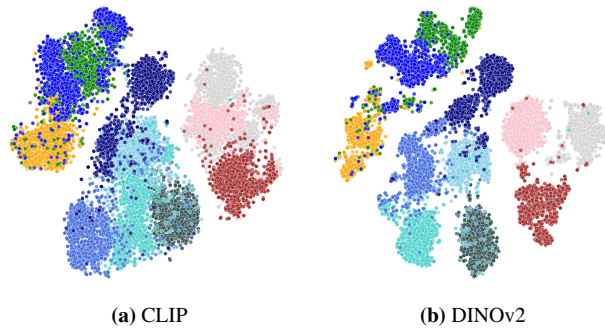


**Figure 2.** Performance of vision foundation models across different OOD splits. The evaluation metric is FPR95, with lower values indicating better performance.

#### 3.2. Evaluation of Vision Foundation Models

Although several studies [11, 40, 41, 70] have explored the use of foundation models for OOD detection, they focus solely on vision-language foundation models such as CLIP [46]. Beyond CLIP, the community offers a variety of vision foundation models that provide robust raw feature space. This development has inspired us to re-examine which vision foundation model is best suited for OOD detection. In this section, we aim to investigate and analyze various pre-trained vision foundation models as effective OOD detectors without fine-tuning.

**Experimental Setup.** We perform our evaluation on a challenging OOD detection benchmark that utilizes ImageNet-1K as ID data and selects samples from iNaturalist18, Sun, Places, and Textures as OOD samples. We also include two challenging OOD test sets: ImageNet-O [7] and NINCO [19]. We choose several representative vision foundation models, namely DINOv1 [2], DINOv2 [38], MAE [13], SigCLIP [75], OpenCLIP [18], and OpenAI CLIP [46]. For fair comparison, we use the ViT-B as the architecture of these models. The scoring functions for DINOv1, DINOv2, and MAE are set to KNN [54]. For the CLIP series, we report the best results among four scoring functions (MSP, MaxLogit, Energy and KNN). Without any model tuning, we directly use the features extracted from these models for OOD detection evaluation to assess whether they are already sufficiently capable of OOD detection. To emphasize the



**Figure 3. Feature Visualization for Foundation Models.** For fine-grained feature visualization, we randomly select fine-grained categories under 3 different super classes from ImageNet-1k.

significance of our findings, we also compare them with the state-of-the-art method (NegPrompt [30]) that involve fine-tuning an ImageNet pre-trained model on the ID dataset. Additionally, we conduct further verification using datasets beyond ImageNet-1k as ID data in Appendix A.4 and the conclusion align with the experiments using ImageNet-1k.

**Result Analysis.** (1) With traditional score metrics (*i.e.* KNN), DINOv2 can outperform all other foundation models by a large margin. DINOv2+KNN shows the best results where the average FPR95 is 29.27% in the first four test sets, 8.9% in the latter challenging test sets, while SigCIIP only achieves 54.23% and 77.17%. This is potentially because DINOv2 leverages advanced self-supervised learning: iBot [80], which is a Mask Image Modeling (MIM) pre-task for facilitating models to capture image details, and contrastive learning objective [2] that enhances the feature discriminativeness. (2) Without any fine-tuning, DINOv2 achieves performance comparable to the more complex method (*i.e.* NegPrompt) in the first four test sets. Notably, DINOv2+KNN still significantly outperforms NegPrompt on challenging test sets by 49.94%. The reason is that these two datasets contain images that are extremely similar to the ID categories. However, the paradigm of CLIP only provides image-level textual supervision without a supervision signal to retain detailed image information. Therefore, CLIP always fails in some fine-grained tasks, while DINOv2 consistently performs much better. As shown in Fig. 3a and Fig. 3b, where we randomly select 11 fine-grain categories under 3 different super classes, DINOv2 provide more discriminative boundaries, while CLIP can not.

**Further Challenges in OOD using Foundation models.** In summary, DINOv2, without requiring any fine-tuning, can already function as a high-performing OOD detector, surpassing previous approaches and underscoring the importance of discriminative and generalizable features for OOD detection. However, foundation models still have room to improve

and cannot generalize well across the entire feature space. (1) Though there is a consensus that fine-tuning on the ID data can improve OOD performance [4, 15, 55, 67], we find that this doesn't hold in the context of foundation models, particularly on in-domain data with large semantic spaces. For instance, when we fine-tune DINOv2 on ImageNet-1K ID data and evaluate the fine-tuned model, the performance declines on three out of the four OOD datasets (see naive finetuning in Tab. 7 in Appendix). The implementation details of this finetuning can be referred to Appendix A.2. (2) Besides, as shown in Fig. 7c and Fig. 7f in Appendix, we also show some failure examples, where the models exhibit particularly poor feature discriminability, hindering effective OOD detection.

## 4. Method

This section introduce our proposed methods for finetuning vision foundation models to enhance the OOD detection ability, which includes a Mixture of Feature Expert module in Sec. 4.1 and a Dynamic- $\beta$  Mixup data augmentation strategy in Sec. 4.2.

### 4.1. Mixture of Feature Experts

As shown in Fig. 4, we propose Mixture of Feature Experts (MoFE), which divides the complex semantic space into multiple subspaces and each expert specializes in a specific subspace. Each expert can tackle an easier problem instead of conducting OOD detection on a complicated distribution, which eases the optimization process while maintaining the generalizability of features. Below presents the detailed configuration of MoFE.

Given an RGB image  $\mathbf{v} \in \mathbb{R}^{H \times W \times 3}$ , where  $H$  and  $W$  are the origin resolution, we reshape the image  $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$  into a sequence of flattened 2D patches  $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 \cdot C)}$ ,  $C$  is the number of channels,  $(P, P)$  is the resolution of each image patch. Next, we flatten the patches and map to  $D$  dimensions with a trainable linear projection  $\mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}$ . A learnable embedding is prepended to the sequence of embedded patches ( $\mathbf{z}_0^0 = \mathbf{x}_{\text{cls}}^0$ ) and position embeddings are added to the patch embeddings  $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1) \times D}$ . Then we input these embeddings to multiple transformer blocks. The output is processed by a MoFE layer to obtain the domain-specific features. This process is expressed as:

$$\mathbf{z}_0 = [\mathbf{x}_{\text{cls}}^0; \mathbf{x}_p^1 \mathbf{E}; \mathbf{x}_p^2 \mathbf{E}; \dots; \mathbf{x}_p^N \mathbf{E}] + \mathbf{E}_{\text{pos}}, \quad (2)$$

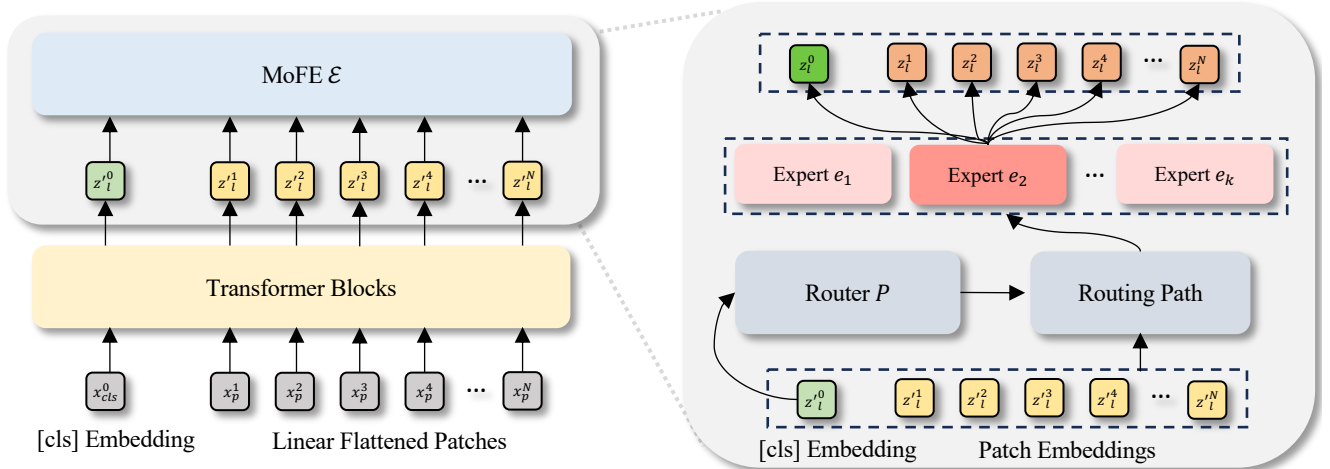
$$\mathbf{z}'_\ell = \text{Transformer}(\mathbf{z}_{\ell-1}) + \mathbf{z}_{\ell-1}, \ell = 1 \dots L, \quad (3)$$

$$\mathbf{z}_\ell = \text{MoFE}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, \quad (4)$$

$$\mathbf{F} = \text{LN}(\mathbf{z}_l). \quad (5)$$

where  $\text{LN}$  denotes the layer norm.

**MoFE Architecture.** The MoFE layer consists of multiple expert networks, each of which is a transformer block. As an



**Figure 4. Illustration of our proposed Mixture of Feature Experts (MoFE).** MoFE decomposes the large semantic space into multiple subspaces and each expert specializes in a specific subspace. Specifically, the image patches and the class token are input to obtain the preliminary patch embeddings and class embedding. A router is employed to determine the expert to further process the embeddings, and the input of the router is the class embedding. Finally, we apply associated experts to refine the class embeddings and the patch embeddings. We use the class embeddings output by MoFE and conduct the OOD detection in the corresponding subspace.

initialization step, we replicate the transformer blocks from the final layer of a foundation model to form an ensemble of experts  $\mathcal{E} = [e_1, e_2, \dots, e_E]$ . The router [50] is a linear layer that predicts the probability of each token being assigned to each expert. Routing accuracy is crucial for MoFE. The key question is what should be used to determine the results of feature routing? We explore various approaches, such as reinitializing a routing token, averaging patch embeddings, or utilizing class embeddings. We ultimately find that using the class embedding achieves the best results. Although it is not the embedding output from the last layer of the network, it is sufficiently discriminative. Therefore, we utilize the class embedding  $\mathbf{z}_l^0$  as the input of the router. The router is a linear layer that predicts the probability of each token being assigned to each expert. We formulate as:

$$\mathcal{P}(\mathbf{z}_l^0)_i = \frac{e^{f(\mathbf{z}_l^0)_i}}{\sum_j^E e^{f(\mathbf{z}_l^0)_j}}, \quad (6)$$

where the router produces weight logits  $f(\mathbf{z}_l^0) = \mathbf{W} \cdot \mathbf{z}_l^0$ , which are normalized by the softmax function.  $\mathbf{W} \in \mathbb{R}^{D \times E}$  represents the lightweight training parameters and  $E$  represents the number of experts. After determining the experts by using the class embedding, we input all embedding including the patch embeddings and class embedding to the activated experts. Each embedding is processed by the top- $k$  experts with the highest probabilities, and the weighted sum is calculated based on the softmax results of the probabilities:

$$\text{MoFE}(\mathbf{z}'_\ell) = \sum_{i=1}^k \mathcal{P}(\mathbf{z}'_\ell)_i \cdot \mathcal{E}(\mathbf{z}'_\ell)_i, \quad (7)$$

where  $\mathcal{E}$  represents the network of an expert [12]. In our

MoFE architecture, we route only a single expert, thus  $k = 1$ . We find that the router computation is reduced as we are only routing a token to a single expert and the performance does not increase when using more experts.

**Feature Space Separation.** In MoFE, we aim to have different experts specialize in different subspaces. Therefore, we propose to first separate the whole feature space into multiple subspaces so that each expert specializes in learning features within its subspace. We use WordNet [39], which provides a good summary of the higher-level semantics of categories, to offer an initial partition of the subspace. Since the semantic and visual similarities are not completely equivalent, we further refine the clustering using K-Means to adjust for the discrepancies. Specifically, we extract feature representations  $\mathbf{z}'_l$  for each training image. Then, we calculate the class prototypes by averaging the features of the images from each category. Finally, we perform a K-Means clustering on categorical feature prototypes. The initial cluster centers are determined by the centroids of clusters, which are divided according to the original semantic space. After determining the clustering results, we assign different experts to different clusters, with samples from each category being routed to the corresponding expert model.

**MoFE Training.** We replace the final transformer block with a MoFE layer. Each transformer block within MoFE is initialized by the original final transformer block. We could set multiple layers as MoFE layers, but we find that using just the final layer achieves sufficiently good results. Then we randomly initialize a router layer and use the class token as the input. We use the labels generated by the above

clustering to supervise the routing:

$$\mathcal{L}_{\text{route}} = -\frac{1}{N} \sum_{n=1}^N \sum_{i=1}^E y^i \log(\mathcal{P}^i(\mathbf{z}'_i^0)). \quad (8)$$

For each expert, we leverage the categories within the corresponding cluster as the positive samples, and the categories beyond the cluster as the negative ones. Assuming that the category cluster of the  $i$ th expert contains  $Q_i$  classes, we set the categories beyond the cluster as the  $Q_i + 1$  categories. The loss is designed as follows:

$$\mathcal{L}_{\text{expert}} = -\frac{1}{N} \sum_{n=1}^N \sum_{i=1}^E \sum_{q=1}^{Q_i+1} y^i y^q \log(p_i^q(\mathbf{x})). \quad (9)$$

In order to achieve the sample balance for each cluster, we control the ratio of positive and negative samples as 1:1 during training. Therefore, the overall loss of MoFE is:

$$\mathcal{L}_{\text{MoFE}} = \mathcal{L}_{\text{expert}} + \mathcal{L}_{\text{route}}. \quad (10)$$

**Discussion.** MoFE is designed to address OOD issues under large-scale complex distributions. Similar to MOS [17], we decouple complex distributions into simpler subspace. However, while MOS approaches this solely from the perspective of the loss function, we approach it from the model perspective by assigning an expert model to each subspace. This allows each expert to focus on learning its assigned subspace, preventing interference between features from different partitions. To accurately assign expert models to different samples during inference, we have devised a new routing method that uses the [CLS] token as the input to the router network, since it encapsulates the semantic features of the entire image. With these designs, MoFE outperforms MOS by a substantial margin (see Tab. 1) and the feature visualization shows that MoFE can achieve more compact ID distribution and clearer decision boundaries between ID and OOD samples (Appendix Fig. 5).

## 4.2. Dynamic- $\beta$ Mixup

Data augmentation (*e.g.*, Mixup [64, 77]) has been proven to improve generalization during finetuning. Traditional Mixup [64, 77] augment samples and transform labels by:

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j, \quad \tilde{y} = \lambda y_i + (1 - \lambda)y_j, \quad (11)$$

where  $\lambda \sim \text{Beta}(\sigma, \sigma)$ .  $\lambda$  is the interpolation weight for generating new augmented samples. We observe that different categories exhibit varying levels of discriminativeness initialized by vision foundation models, as shown in Appendix Fig. 7c and Fig. 7f. For categories that are already well-represented, synthesizing dissimilar samples via vanilla Mixup can blur the decision boundary between ID

and OOD, leading to degraded performance (Fig. 4 in Appendix). Therefore, we dynamically adjust the Beta distribution according to the feature discriminativeness per category. The reason is that when features of  $x_i$  are discriminative enough, a small  $\lambda$ , which leads to a dissimilar sample, is not necessary for their representation learning. Instead, we should leverage similar samples from a large  $\lambda$  for building smooth decision boundaries. On the contrary, when features of a category show poor discriminativeness, we should set a relatively small  $\lambda$  to ease the feature learning. We use the accuracy of the validation set to measure the discriminativeness. Therefore, we set  $\lambda$  as:

$$\lambda \sim \text{Beta}(\sigma, \sigma) \text{ for } \sigma = 1 - w * s, \quad (12)$$

where  $w$  is a scaling factor and  $s$  denotes the corresponding category’s accuracy on the validation set. Because the probability density function of  $\text{Beta}(\sigma, \sigma)$  is symmetric about 0.5 and ranges from 0 to 1, we need to ensure that with a larger  $s$ , the probability of sampling larger values is greater. Therefore, we transform  $\lambda$  as:

$$\hat{\lambda} = \begin{cases} \lambda & \lambda \geq 0.5 \\ 1 - \lambda & \lambda < 0.5 \end{cases}. \quad (13)$$

We determine the category difficulty at the beginning of the training and then update it during the training process. In our implementation,  $x_i$  is the training sample, and  $x_j$  is the instance used to corrupt  $x_i$ . Therefore, we select the  $s$  from categories of  $x_i$ , and we select samples from different classes. Additionally, we empirically find that using vanilla Mixup [64, 77] can cause feature norms to grow during finetuning vision foundation models (*i.e.*, DINOv2), leading to performance degradation on the OOD task. In order to restrain the growth of feature norms, we propose to add a regularization term to suppress the increase in feature norm:

$$\mathcal{L}_{\text{Mixup}} = -\frac{1}{N} \sum_{n=1}^N \sum_{c=1}^C y^c \log(p^c(x)) + \text{Reg}(F^0), \quad (14)$$

where  $C$  is the total number of categories,  $\text{Reg}$  denotes a regularization method,  $F^0$  is the final class embeddings output by MoFE. By default, the regularization method has multiple choices, which can be  $L_2$  norm or label smoothing. The final optimization objective is:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{MoFE}} + \mathcal{L}_{\text{Mixup}}. \quad (15)$$

## 5. Experiments

In this section, we set up a benchmark for evaluating OOD performance in Sec. 5.1. Then we compare our methods with the competitive baselines in Sec. 5.2. We conduct ablation studies and present more analysis on our designed method in Sec. 5.3.

Method	iNaturalist18		Places		Sun		Textures		Average		ID ACC $\uparrow$	
	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$		
CLIP-Based	Energy [35]	65.00	87.17	57.40	87.32	46.43	91.17	57.40	87.32	56.55	88.24	79.39
	MSP [14]	40.89	88.63	65.81	81.24	67.90	80.14	64.96	78.16	59.89	82.04	79.39
	MaxLogit [15]	60.86	88.03	55.5	87.44	44.81	91.16	52.25	86.04	53.35	88.16	<b>79.39</b>
	MCM [41]	30.91	94.61	37.59	92.57	44.69	89.77	57.77	86.11	42.74	90.77	67.01
	CLIPN [70]	23.94	95.27	26.17	93.93	33.45	92.28	40.83	90.93	31.10	93.10	68.53
	LSN [45]	21.56	95.83	34.48	91.25	26.32	94.35	38.54	90.42	30.22	92.96	71.89
	NegPrompt [30]	6.32	<b>98.43</b>	27.60	93.34	22.89	<b>95.55</b>	35.21	91.60	23.01	94.81	66.84
<i>Ours</i>	<b>5.19</b>	97.28	<b>21.32</b>	<b>94.69</b>	<b>22.10</b>	95.17	<b>31.47</b>	<b>92.15</b>	<b>20.02</b>	<b>94.89</b>	68.56	
Dinov2-Based	MSP [14]	25.02	94.76	57.09	83.45	53.65	85.22	48.79	85.81	48.13	87.31	86.01
	MaxLogit [15]	22.96	94.59	59.21	78.41	54.52	81.80	48.17	84.16	46.21	84.74	86.01
	Energy [35]	28.48	93.19	65.88	74.49	61.54	78.71	53.29	81.92	52.29	82.07	86.01
	KNN [54]	5.67	97.65	43.25	88.21	36.42	90.21	28.04	92.66	28.34	92.18	86.01
	MOS [17]	5.01	97.85	40.15	90.33	34.32	91.87	26.14	92.98	26.40	93.25	85.23
	<i>Ours</i>	<b>2.74</b>	<b>98.82</b>	<b>24.32</b>	<b>93.73</b>	<b>17.38</b>	<b>95.65</b>	<b>18.58</b>	<b>95.38</b>	<b>17.01</b>	<b>95.89</b>	<b>86.40</b>

**Table 1. Quantitative results of OOD detection performance for ImageNet-1k as ID.** We employed our method on two pre-training paradigms (CLIP, and DINOv2). We use FPR95 and AUROC as evaluation metrics. We also report ID classification accuracy. The CLIP-based methods use ViT-B-16, and the DINOv2-based methods use ViT-B-14.

## 5.1. Benchmark

**In- and out-distribution Datasets.** To validate the effectiveness of our proposed method, we conduct evaluation on standard benchmarks, which use ImageNet-1K [48] and ImageNet-100 [41] as the ID datasets. Following existing studies that leverage foundation models in OOD [30, 70], we use diverse OOD test datasets, including samples selected from iNaturalist18 [66], SUN [74], Places [79], and Textures [6].

**Method Comparison.** We conduct method comparison on two pretraining paradigms (*i.e.* CLIP and DINOv2). For each group, we apply some traditional scoring metric (such as MSP [14], MaxLogit [15], Energy [35], KNN [54]). Moreover, we also involve the current CLIP-based state-of-the-art methods, such as MCM [41], CLIPN [70], NegPrompt [30], and LSN [45]. We use KNN [54] as the scoring function when using DINOv2, and follow the scoring metric of CLIPN [70] when applying our method to CLIP.

## 5.2. Main Results

**Results on ImageNet-1K.** We compare the proposed approach with the state-of-the-art methods for ImageNet-1K as ID on Tab. 1. These results show: 1) Based on DINOv2, our method reaches the best performance when setting ImageNet-1K as ID. Specifically, our approach reaches 17.01% FPR95 and 95.89% AUROC, averaging the results of all the OOD test sets. Our method surpasses MOS [17] by 9.39% in FPR95, and 2.64% in AUROC, which proves the importance of learning expert models for OOD detection. 2) When applying our method to CLIP, our method reaches 20.02% and 94.89%, which also outperforms NegPrompt [30] by

a large margin. These results indicate the effectiveness of the proposed MoFE and the dynamic regularized Mixup. 3) Our approach reaches 2.74% FPR95 on iNaturalist18 and increases the performance on all the test sets, which indicates that our MoFE design retains the discriminativeness of DINOv2 and facilitates feature learning on various feature subspaces.

**Results on ImageNet-100.** We compare the proposed approach with the state-of-the-art methods for ImageNet-100 as ID on Tab. 2. Based on DINOv2, our method reaches 8.10% FPR95 and 97.75% AUROC, surpassing the baseline by 4.40% FPR95, and 0.23% AUROC. This indicates that our proposed approach is also effective in a small-scale ID dataset. On the other hand, when applying our method to CLIP, we achieve 7.40% FPR95 and 98.10% AUROC, outperforming LSN [45] by 1.16% FPR95. The above experimental results validate the effectiveness of our approach, and we achieve the best performance on both small-scale and large-scale ID datasets.

## 5.3. Analysis

In this section, we conduct more analysis on the proposed methods. We use the DINOv2 as the pretraining paradigm and KNN [54] as the scoring function. We use ImageNet-1K as ID data and report the average performance on the four out-of-distribution datasets mentioned in Sec. 5.1. The baseline is set to DINOv2 with KNN scoring function without our designed method.

**Contributions of Individual Components.** As shown in Tab. 3, we evaluate the contribution of each component of the full method. On ImageNet-1K, MoFE and Dynamic- $\beta$  Mixup contribute 6.68% and 5.41% FPR95, respectively.

Method	iNaturalist18		Places		Sun		Textures		Average		ID ACC $\uparrow$	
	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$	FPR95 $\downarrow$	AUROC $\uparrow$		
CLIP-Based	MSP [14]	23.55	95.92	40.46	91.23	37.02	92.45	24.40	94.90	31.43	93.63	91.93
	MCM [41]	18.13	96.77	34.52	94.36	36.45	94.54	41.22	92.25	32.58	94.48	87.88
	CLIPN [70]	4.87	98.16	13.64	96.93	13.55	97.56	15.78	93.02	11.96	96.41	91.64
	LSN [45]	4.93	98.92	12.82	97.19	8.23	97.98	<b>8.26</b>	<b>98.11</b>	8.56	98.05	92.24
	<i>Ours</i>	<b>3.20</b>	<b>99.17</b>	<b>10.05</b>	<b>97.76</b>	<b>7.06</b>	<b>98.39</b>	9.31	97.10	<b>7.40</b>	<b>98.10</b>	<b>92.85</b>
Dinov2-Based	MSP [14]	5.06	98.85	26.58	94.78	27.64	95.02	26.43	94.27	21.42	95.72	94.50
	MaxLogit [15]	5.55	98.76	29.69	94.19	32.73	94.20	29.27	93.72	24.31	95.21	94.50
	Energy [35]	18.57	96.69	54.72	88.92	62.42	87.17	57.28	88.40	48.21	90.29	94.50
	KNN [54]	2.58	99.02	18.45	95.12	15.89	96.16	16.79	96.38	13.42	96.66	94.50
	<i>Ours</i>	<b>2.25</b>	<b>99.23</b>	<b>12.81</b>	<b>96.66</b>	<b>8.51</b>	<b>97.86</b>	<b>8.85</b>	<b>97.28</b>	<b>8.10</b>	<b>97.75</b>	<b>96.94</b>

Table 2. Quantitative results of OOD detection performance for ImageNet-100 as ID.

Settings	Baseline	+ MoFE	+ D- $\beta$	+ MoFE+D- $\beta$
FPR95 $\downarrow$	29.27	22.59	23.85	<b>17.01</b>
AUROC $\uparrow$	92.67	94.01	93.72	<b>95.89</b>

Table 3. Ablation study of individual components for ImageNet-1k as In-Distribution dataset.

Num.	2	3	5	7	8	9
Gain	4.10	6.30	9.80	12.26	12.10	12.09

Table 4. The effect of Cluster Number. We report the performance gain in FPR95 compared to the model without MoFE.

Grouping	Baseline	WordNet	Clustering	Ours
FPR95 $\downarrow$	29.27	25.63	26.34	<b>22.59</b>
AUROC $\uparrow$	92.67	93.01	92.99	<b>94.01</b>

Table 5. Analysis on Grouping Strategy in MoFE.

When combined, the best performance (17.01% FPR95, and 95.89% AUROC) is achieved.

**Cluster Number.** We conduct an experiment to validate the impact of cluster number on MoFE performance. We set different numbers of clusters. As shown in Tab. 4, we report the performance gain in FPR95. As the increasing of cluster number, the performance gradually increases.

**Feature Space Separation.** As shown in Tab. 5, we validate the different strategies for determining the subspace. We compare our method with two methods: WordNet and clustering. The results show that using the ours (i.e. WordNet + Clustering) is the most promising approach. The reason might be that the features extracted by pretrained model are discriminative, especially at coarse-grained level.

**More Analysis on Dynamic- $\beta$  Mixup.** As shown in Tab. 6, we conduct an ablation study on Dynamic- $\beta$  Mixup. When we remove the regularization term, we find that the performance degrades (30.43% FPR95). Moreover, when we dynamic beta distribution is removed, the performance decreases to 24.96% FPR95.

Methods	Baseline	w/o Reg	w/o D- $\beta$	Ours
FPR95 $\downarrow$	29.27	30.43	24.96	<b>23.85</b>
AUROC $\uparrow$	92.67	91.65	93.36	<b>93.72</b>

Table 6. Ablation study of Dynamic- $\beta$  Mixup.

## 6. Conclusion

This paper studies the OOD detection task within the context of foundation models. Our study shows that vision foundation models (e.g., DINOv2) are effective OOD detectors, suggesting high-quality and generalizable feature space is essential for OOD detection. Our study highlights that CLIP’s pre-trained feature space is less effective for fine-grained tasks, where DINOv2 performs significantly better, which worths further exploration. In order to further optimize the performance of OOD detection when ID data is available for fine-tuning, we propose MoFE and a Dynamic- $\beta$  Mixup data augmentation to enhance the feature learning. We conduct extensive experiments and ablation studies to validate the effectiveness of our approach. We hope our investigation could inspire more future studies.

## Acknowledgments

This research is supported by the National Key RD Program of China (Grant No. 2023YFB2806300). This work has been supported in part by Hong Kong Research Grant Council - Early Career Scheme (Grant No. 27209621), General Research Fund Scheme (Grant No. 172024221721292317215025), Theme-based Research (Grant No. T45-701/22-R) and Shenzhen Science and Technology Innovation Commission (SGDX20220530111405040). Part of the described research work is conducted in the JC STEM Lab of Robotics for Soft Materials funded by The Hong Kong Jockey Club Charities Trust.

## References

- [1] Sparse mixture-of-experts are domain generalizable learners, author=Li, Bo and Shen, Yifei and Yang, Jingkang and Wang, Yezhen and Ren, Jiawei and Che, Tong and Zhang, Jun and Liu, Ziwei, booktitle=ICLR, year=2023. 1
- [2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 3, 4
- [3] Chirui Chang, Zhengzhe Liu, Xiaoyang Lyu, and Xiaojuan Qi. What matters in detecting ai-generated videos like sora? *arXiv preprint arXiv:2406.19568*, 2024. 2
- [4] Guangyao Chen, Limeng Qiao, Yemin Shi, Peixi Peng, Jia Li, Tiejun Huang, Shiliang Pu, and Yonghong Tian. Learning open set network with discriminative reciprocal points. In *ECCV*, 2020. 4
- [5] Shaoxiang Chen, Zequn Jie, and Lin Ma. Llava-mole: Sparse mixture of lora experts for mitigating data conflicts in instruction finetuning mllms. *arXiv preprint arXiv:2401.16160*, 2024. 3
- [6] Mircea Cimpoi, Subhansu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 7
- [7] Hendrycks Dan, Zhao Kevin, Basart Steven, Steinhardt Jacob, and Song Dawn. Natural adversarial examples. In *CVPR*, 2021. 3
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1
- [9] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. In *ICLR*, 2022. 1, 2
- [10] Xuefeng Du, Yiyu Sun, Xiaojin Zhu, and Yixuan Li. Dream the impossible: Outlier imagination with diffusion models. In *NeurIPS*, 2023. 1
- [11] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. Zero-shot out-of-distribution detection based on the pre-trained model clip. In *AAAI*, 2022. 1, 2, 3
- [12] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *CoRR*, 2021. 2, 3, 5
- [13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *CVPR*, 2022. 3
- [14] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017. 1, 2, 3, 7, 8
- [15] 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, 4, 7, 8
- [16] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *CVPR*, 2020. 1
- [17] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *CVPR*, 2021. 2, 6, 7, 3
- [18] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 3
- [19] Bitterwolf Julian, Mueller Maximilian, and Matthias. Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In *ICML*, 2023. 3
- [20] Umar Khalid, Ashkan Esmaeili, Nazmul Karim, and Nazanin Rahnavard. Rodd: A self-supervised approach for robust out-of-distribution detection. In *CVPRW*, 2022. 2
- [21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 1
- [22] Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. Sparse upcycling: Training mixture-of-experts from dense checkpoints. *arXiv preprint arXiv:2212.05055*, 2022. 3
- [23] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from <https://github.com/openimages>*, 2017. 1, 3
- [24] Yamuna Krishnamurthy, Chris Watkins, and Thomas Gaertner. Improving expert specialization in mixture of experts. *arXiv preprint arXiv:2302.14703*, 2023. 3
- [25] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *ICLR*, 2018. 1
- [26] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018. 1, 2
- [27] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020. 3
- [28] Jinglun Li, Xinyu Zhou, Pinxue Guo, Yixuan Sun, Yiwen Huang, Weifeng Ge, and Wenqiang Zhang. Hierarchical visual categories modeling: A joint representation learning and density estimation framework for out-of-distribution detection. In *ICCV*, 2023. 2
- [29] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 1

- [30] Tianqi Li, Guansong Pang, Xiao Bai, Wenjun Miao, and Jin Zheng. Learning transferable negative prompts for out-of-distribution detection. In *CVPR*, 2024. 4, 7
- [31] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018. 1
- [32] Randolph Linderman, Jingyang Zhang, Nathan Inkawich, Hai Li, and Yiran Chen. Fine-grain inference on out-of-distribution data with hierarchical classification. *arXiv preprint arXiv:2209.04493*, 2022. 2
- [33] Jiahui Liu, Chirui Chang, Jianhui Liu, Xiaoyang Wu, Lan Ma, and Xiaojuan Qi. Mars3d: A plug-and-play motion-aware model for semantic segmentation on multi-scan 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9372–9381, 2023. 2
- [34] Jiahui Liu, Xin Wen, Shizhen Zhao, Yingxian Chen, and Xiaojuan Qi. Can ood object detectors learn from foundation models? In *European Conference on Computer Vision*, pages 213–231. Springer, 2024. 2
- [35] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *NeurIPS*, 2020. 1, 2, 3, 7, 8
- [36] Yuxuan Lou, Fuzhao Xue, Zangwei Zheng, and Yang You. Cross-token modeling with conditional computation. *arXiv preprint arXiv:2109.02008*, 2021. 3
- [37] Xiaoyang Lyu, Chirui Chang, Peng Dai, Yang-Tian Sun, and Xiaojuan Qi. Total-decom: decomposed 3d scene reconstruction with minimal interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20860–20869, 2024. 2
- [38] Oquab Maxime, Darcet Timothée, Moutakanni Théo, Vo Huy, Szafraniec Marc, Khalidov Vasil, Fernandez Pierre, Haziza Daniel, Massa Francisco, El-Nouby Alaaeldin, Assran Mahmoud, Ballas Nicolas, Galuba Wojciech, Howes Russell, Huang Po-Yao, Li Shang-Wen, Misra Ishan, Rabbat Michael, Sharma Vasu, Synnaeve Gabriel, Xu Hu, Jegou Hervé, Mairal Julien, Labatut Patrick, Joulin Armand, and Bojanowski Piotr. Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193*, 2023. 1, 2, 3
- [39] George A. Miller. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994. 5
- [40] Yifei Ming and Yixuan Li. How does fine-tuning impact out-of-distribution detection for vision-language models? *IJCV*, 2024. 1, 2, 3
- [41] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyu Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. In *NeurIPS*, 2022. 1, 3, 7, 8, 2
- [42] Atsuyuki Miyai, Qing Yu, Go Irie, and Kiyoharu Aizawa. Locoop: Few-shot out-of-distribution detection via prompt learning. In *NIPS*, 2023. 2
- [43] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *arXiv preprint arXiv:2206.02770*, 2022. 3
- [44] 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. 2
- [45] Jun Nie, Yonggang Zhang, Zhen Fang, Tongliang Liu, Bo Han, and Xinmei Tian. Out-of-distribution detection with negative prompts. In *ICLR*, 2024. 2, 3, 7, 8
- [46] 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 *ICML*, 2021. 1, 2, 3
- [47] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *NeurIPS*, 2021. 3, 1
- [48] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015. 7
- [49] Vikash Sehwal, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. In *ICLR*, 2021. 1, 2
- [50] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017. 2, 3, 5
- [51] Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollár, Christoph Feichtenhofer, Ross Girshick, Rohit Girdhar, and Ishan Misra. The effectiveness of mae pre-training for billion-scale pretraining. In *ICCV*, 2023. 1
- [52] Yiyu Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *ECCV*, 2022. 1
- [53] Yiyu Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *NeurIPS*, 2021. 1
- [54] Yiyu Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *ICML*, 2022. 1, 2, 3, 7, 8
- [55] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *NeurIPS*, 2020. 2, 4
- [56] Haoru Tan, Chuang Wang, Sitong Wu, Tieqiang Wang, Xu-Yao Zhang, and Cheng-Lin Liu. Proxy graph matching with proximal matching networks. In *The Annual AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 2
- [57] Haoru Tan, Sitong Wu, Fei Du, Yukang Chen, Zhibin Wang, Fan Wang, and Xiaojuan Qi. Data pruning via moving-one-sample-out. In *Neural Information Processing Systems (NeurIPS)*, 2023. 2
- [58] Haoru Tan, Sitong Wu, and Jimin Pi. Semantic diffusion network for semantic segmentation. In *Neural Information Processing Systems (NeurIPS)*, 2023.

- [59] Haoru Tan, Chuang Wang, Sitong Wu, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. Ensemble quadratic assignment network for graph matching. *International Journal of Computer Vision (IJCV)*, 2024.
- [60] Haoru Tan, Sitong Wu, Zhuotao Tian, Yukang Chen, Xiaojuan Qi, and Jiaya Jia. Saco loss: Sample-wise affinity consistency for vision-language pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [61] Haoru Tan, Sitong Wu, Wei Huang, Shizhen Zhao, and Xiaojuan Qi. Data pruning by information maximization. In *International Conference on Learning Representations (ICLR)*, 2025. 2
- [62] Haoru Tan, Sitong Wu, Bo Zhao, Zeke Xie, and XIAOJUAN QI. Diff-in: Data influence estimation with differential approximation, 2025. 2
- [63] Leitian Tao, Xuefeng Du, Xiaojin Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *ICLR*, 2023. 1
- [64] Sunil Thulasidasan, Gopinath Chennupati, Jeff Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *NeurIPS*, 2019. 2, 6
- [65] Changyao Tian, Wenhai Wang, Xizhou Zhu, Xiaogang Wang, Jifeng Dai, and Yu Qiao. V1-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition. *arXiv preprint arXiv:2111.13579*, 2021. 1
- [66] Grant Van Horn, Oisin 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 *CVPR*, 2018. 7
- [67] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: a good closed-set classifier is all you need? In *ICLR*, 2022. 4
- [68] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *CVPR*, 2022. 1
- [69] Haotao Wang, Aston Zhang, Yi Zhu, Shuai Zheng, Mu Li, Alex J Smola, and Zhangyang Wang. Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition. In *ICML*, 2022. 2
- [70] Hualiang Wang, Yi Li, Huifeng Yao, and Xiaomeng Li. Clipn for zero-shot ood detection: Teaching clip to say no. In *ICCV*, 2023. 1, 2, 3, 7, 8
- [71] Hai Wu, Ruifei He, Haoru Tan, Xiaojuan Qi, and Kaibin Huang. Vertical layering of quantized neural networks for heterogeneous inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2
- [72] Sitong Wu, Tianyi Wu, Haoru Tan, and Guodong Guo. Pale transformer: A general vision transformer backbone with pale-shaped attention. In *The Annual AAAI Conference on Artificial Intelligence (AAAI)*, 2022. 2
- [73] Sitong Wu, Haoru Tan, Yukang Chen, Shaofeng Zhang, Jingyao Li, Bei Yu, Xiaojuan Qi, and Jiaya Jia. Mixture-of-scores: Robust image-text data quality score via three lines of code. In *International Conference on Computer Vision (ICCV)*, 2025. 3
- [74] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 7
- [75] Zhai Xiaohua, Mustafa Basil, Kolesnikov Alexander, and Beyer Lucas. Sigmoid loss for language image pre-training, 2023. 3
- [76] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 1
- [77] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2017. 6
- [78] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *arXiv preprint arXiv:2206.05836*, 2022. 1
- [79] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *PAMI*, 2017. 7
- [80] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *ICLR*, 2022. 4