A. Appendix

A.1. Implementation Details

We adopt ViT-Base [8] as the backbone. When using pre-training paradigms of CLIP and DINOv2, we directly initialize ViT from their weights. Besides, when using CLIP, we leverage CLIPN [70] as the baseline method and we follow their scoring metric. For DINOv2, we use DINOv2 with standard cross-entropy loss as the baseline method and the scoring metric is KNN [54]. When using DINOv2, we first conduct linear probing for 3 epoches to ensure its training stability. Our models are trained with AdamW optimizer with $\beta_s = \{0.9, 0.95\}$, with an effective batch size of 1024 on 8 NVIDIA 3090 GPUs. The values for weight decay and layer decay are 0.05 and 0.75, The training epochs are set to 40. We set a cosine learning rate schedule and the minimum learning rate is 1e-6.

A.2. Implementation Details of Naive Finetuning

The model is trained with cross entropy loss and Adam optimizer with $\beta_s = \{0.9, 0.95\}$, with an effective batch size of 1024 on 8 NVIDIA 3090 GPUs. The values for weight decay and layer decay are 0.05 and 0.75. The training epochs are set to 40. We set a cosine learning rate schedule, and the minimum learning rate is 1e-6. We first conduct linear probing for 3 epochs to ensure their training stability. During the testing phase, we use KNN as the classifier using features from the penultimate layer.

A.3. Comparison with the traditional MoE

The proposed Mixture of Feature Expert (MoFE) is specifically tailored for OOD detection with foundation models, which is different from the original MoE designed for general LLM and vision tasks from both insights and methods. In terms of insights, our MoFE was crafted to reduces the difficulty of fitting complex data distribution when training foundation models on limited In-Distribution (ID) data, while MoE is initially designed to accelerate inference for large models [47] and is leveraged for learning visual attributes for domain generalization [1]. We're not aware of any existing work that shares our insights. In terms of method design, as our primary insights are to prevent features from collapsing to the ID data distribution, we partition the feature space into different subspaces and design routing mechanism based on feature similarities. Our routing mechanism leverages the class token, which contains the most discriminative feature, to guide all the features to the specific expert.

A.4. Further Evaluation for Pilot Study.

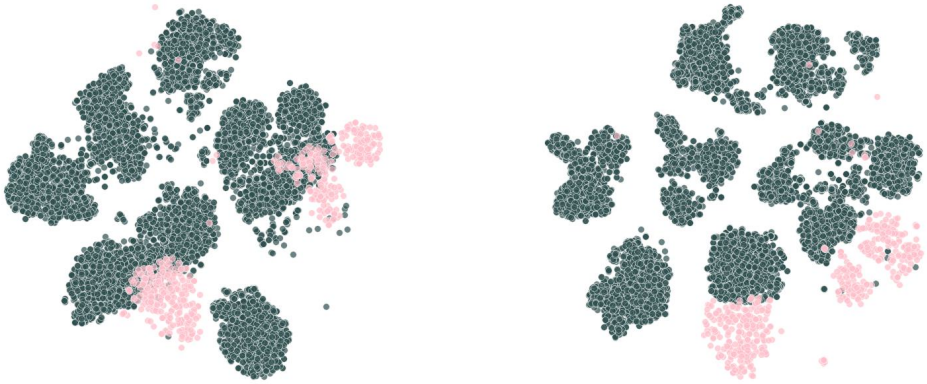
We conduct further validation for pilot study, where we select data from OpenImage [23] for experiments. Specifically, we randomly select 1000 classes as the ID data. Furthermore, we randomly sample another 1000 classes as the OOD data, which is denoted as subset 1. For constructing a finegrained OOD subset, we select the categories which are closely related to the ID categories, where semantically belong to the same superclasses with the ID data according to WordNet. We denote it as subset 2. The results in Tab. 8 demonstrate that 1) Foundation models surpass the ImageNet pretrained methods by a large margin. 2) DINOv2 performs better than CLIP in the finegrained OOD tasks. For example, DINOV2 with KNN achieve 17.23% FPR95 in subset 2, while the CLIP based method can only achieve 29.87% FPR95.

A.5. Limitation

We summarize the limitations of our research as follows: Although CLIP and DINOv2 are currently the top foundation models, they still have inherent shortcomings. For instance, CLIP only utilizes image-text pairs for contrastive learning between text and images, lacking self-supervised learning on images. This results in its inability to capture fine-grained image details, leading to poor performance on finegrained tasks. On the other hand, DINOv2 employs a large number of images for self-supervised learning, yet it still performs poorly on certain categories, indicating potential long-tail distribution issues in its pre-training data. The current benchmarks for OOD detection have significant limitations. While they utilize datasets like ImageNet-1K, which cover a wide range of categories, the OOD data itself is relatively limited.

	Method	iNaturalist18		Places		Sun		Textures		Average		ID ACC↑
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	
IN-1K Pretrained (SoTA)	Energy [35]	55.72	89.95	59.26	85.89	64.92	82.86	53.72	85.99	58.41	86.17	75.08
	MSP [14]	54.99	87.74	70.83	80.86	73.99	79.76	68.00	79.61	66.95	81.99	75.08
	MaxLogit [15]	54.05	87.43	72.98	78.03	73.37	78.03	68.85	79.06	67.31	80.64	75.08
	KNN [54]	7.30	98.46	48.40	88.24	56.46	88.14	39.91	89.23	38.02	91.01	75.08
	MOS [17]	9.54	98.23	43.62	91.26	48.15	90.42	57.12	83.16	39.60	90.76	75.20
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	79.39
	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
Dinov2-Based	Energy [35]	13.23	96.86	66.63	83.32	61.57	84.76	66.43	82.36	51.96	86.82	81.70
	MSP [14]	9.05	98.15	52.58	86.34	49.45	87.35	52.32	85.82	40.85	89.41	81.70
	MaxLogit [15]	8.21	98.22	53.93	85.80	50.48	87.00	54.32	85.25	41.73	89.06	81.70
	KNN [54]	3.01	98.26	42.78	88.89	35.96	91.51	35.30	91.05	29.27	92.67	81.70
	Naive finetuning	5.67	97.65	43.25	88.21	36.42	90.21	28.04	92.66	28.34	92.18	85.96

Table 7. Quantitative results of OOD detection performance for ImageNet-1k as ID. We conduct three pre-training paradigms (ImageNet Pretrained (IN-1K), CLIP, and DINOv2) for comparison. We use FPR95 and AUROC as evaluation metrics. We also report ID classification accuracy.

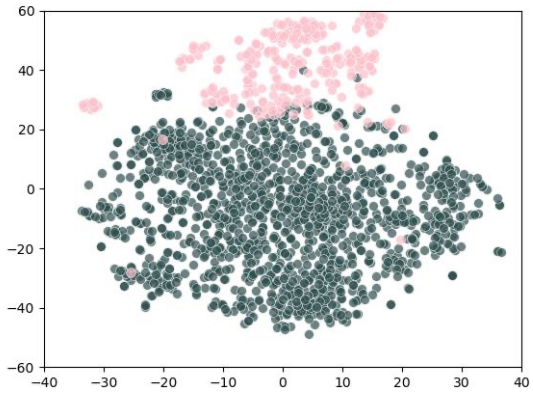


(a) Train with MOS **(b)** Train with MoFE

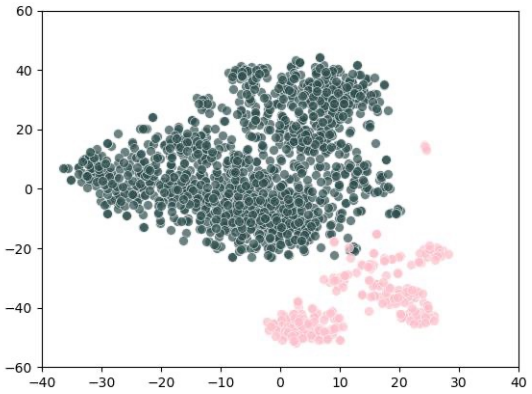
Figure 5. Visualization of feature space of MoFE and MOS. It can be observe that, trained with MOS, the outlier features are still mingled with in-domain data, while MoFE can well separate the in- and out-of-distribution data.

	Method	Subset 1		Subset 2		Average		ID ACC↑
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	AUROC↑	FPR95↓	AUROC↑
IN-1K Pretrained (SoTA)	Energy [35]	60.23	76.23	74.66	73.21	67.44	74.71	72.33
	MSP [14]	58.23	79.01	72.41	77.23	65.32	78.12	72.33
	MaxLogit [15]	57.35	79.32	70.23	78.33	63.79	78.82	72.33
	KNN [54]	15.01	96.55	33.24	94.01	24.12	95.28	72.33
	MOS [17]	17.37	97.01	35.44	93.26	26.41	95.14	73.46
CLIP-Based	Energy [35]	57.43	92.88	65.12	79.23	61.27	86.10	78.64
	MSP [14]	43.23	89.88	62.21	79.11	52.72	84.50	78.64
	MaxLogit [15]	45.87	90.16	60.23	80.12	53.04	85.14	78.64
	MCM [41]	23.34	94.41	45.01	92.16	34.17	93.28	65.27
	CLIPN [70]	10.14	96.88	30.21	94.01	20.18	95.44	64.34
	LSN [45]	9.87	95.12	29.87	95.76	19.87	95.43	72.81
Dinov2-Based	Energy [35]	50.23	88.23	64.13	83.21	57.18	85.71	82.41
	MSP [14]	31.38	93.98	54.32	86.98	42.85	90.48	82.41
	MaxLogit [15]	30.23	94.02	56.32	86.45	43.27	90.23	82.41
	KNN [54]	**8.16**	**97.26**	**17.23**	**96.38**	**12.70**	**96.82**	**82.41**

Table 8. Pilot Study using data from OpenImage [23]. We conduct three pre-training paradigms (ImageNet-1K (IN-1K) Pretrained, CLIP, and DINOv2) for comparison. We use FPR95 and AUROC as evaluation metrics. We also report ID classification accuracy.

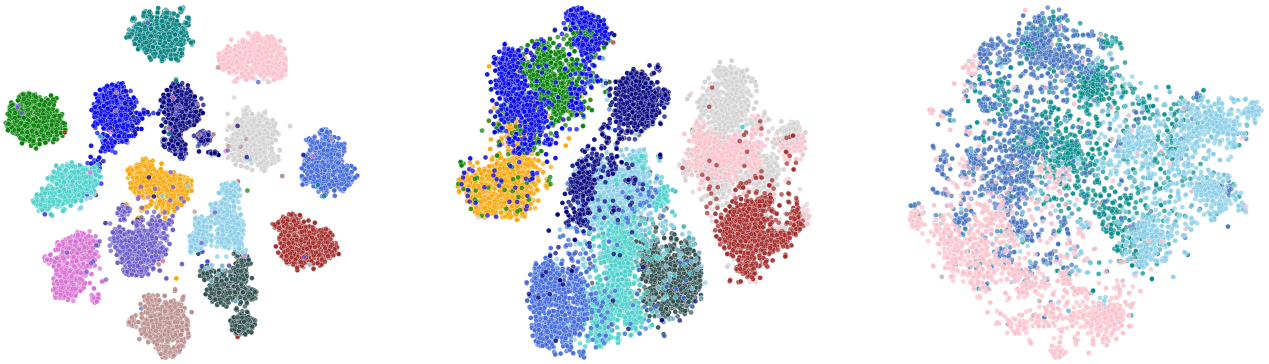


(a) Train with Mixup

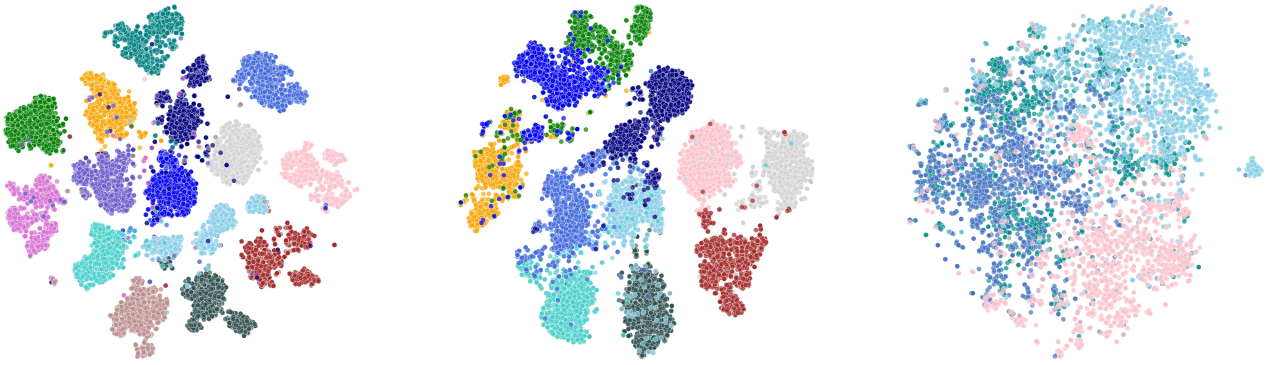


(b) Train without Mixup

Figure 6. The effect of vanilla mixup on the feature space of DINOv2. We can observe that vanilla Mixup can blur the decision boundary between ID and OOD.



(a) CLIP - coarse-grained **(b)** CLIP - fine-grained **(c)** CLIP - failure



(d) DINOv2 - coarse-grained **(e)** DINOv2 - fine-grained **(f)** DINOv2 - failure

Figure 7. Feature Space Visualization for Foundation Models. The first row shows the feature space for CLIP and the second is for DINOv2. For each of them, we visualize the features of coarse-grained categories, fine-grained categories, and some failure cases. For the coarse-grained feature visualization (column 1), we randomly select 15 categories from different super classes in ImageNet-1k following WordNet. For the fine-grained feature visualization (column 2), we randomly select 11 fine-grain categories under 3 different super classes. For the failure case visualization, we select the categories which have the low in-domain accuracy.