Rethinking Out-of-Distribution Detection through the Lens of Model Generalization

Atik Garg[1] Yu-Shuen Wang[2]
[1]EECS International Graduate Program, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
[2]Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
gargatik@gmail.com yushuen@cs.nycu.edu.tw

***Abstract*—We posit that a model's ability to detect out-of-distribution (OOD) samples arises from learning a clear and consistent classification logic. This logic does not need to originate from ground-truth (GT) labels; any coherent, learnable rule derived from the data's inherent structure can suffice. The decisive factor for OOD performance is how well the model generalizes this logic. Our key observation reveals that the performance ceiling in challenging OOD tasks is not solely due to distribution shift, but is fundamentally constrained by the model's failure to generalize on the in-distribution (ID) task itself. Guided by this principle, we adopt an unsupervised framework that trains classifiers on pseudo-labels to form a learnable classification logic. A central feature of our approach is using the model's generalization ability on the ID task as a direct criterion for hyperparameter selection. Experiments demonstrate that this generalization-driven strategy consistently outperforms existing unsupervised methods, highlighting a strong and direct link between generalization and effective OOD detection.**

***Index Terms*—Out-of-distribution, model generalization, pseudo labels**

I. Introduction

In real-world applications, machine learning models inevitably encounter out-of-distribution (OOD) data that differ from the classes observed during training. Although a model may perform well on in-distribution (ID) data, detecting and properly handling OOD samples is crucial for ensuring its trustworthiness. Effective OOD detection mechanisms not only prevent erroneous predictions in unfamiliar scenarios but also enhance model reliability across diverse applications.

Existing OOD detection methods can be broadly categorized into supervised and unsupervised approaches. Supervised methods learn a discriminative classifier over labeled ID classes and then use the resulting scores or representations to identify deviations from the learned distribution as potential OOD samples. These methods are known for their high accuracy and efficiency, but they rely on human-annotated labels, which can be a major limitation in domains where labeling is costly. In contrast, unsupervised methods leverage generative models [1], [2] and detect OOD samples via density estimation [3] or reconstruction-based criteria [4]. While such methods are well-suited for label-scarce scenarios, they frequently struggle with computational efficiency and scalability.

In this study, we argue that supervised methods excel at OOD detection not because they truly recognize real-world semantic concepts (e.g., "car" or "person"), but because they learn a classification logic that is highly specific to the ID dataset. Our central hypothesis is that a model's OOD detection performance is directly correlated with its generalization capability on the ID task. If a model can learn a robust decision boundary that reliably separates different ID classes (i.e., evidenced by ID test accuracy), this same decision logic can be used to identify samples that do not conform to any of the learned classes. From this perspective, supervised methods do not strictly require ground-truth (GT) labels; instead, they primarily require consistent labels that induce a meaningful decision boundary, which can be provided by pseudo-labels.

To validate this hypothesis in a label-free setting, we propose a framework that distills classification logic directly from unlabeled data. We first generate pseudo-labels using self-supervised learning (SSL) and clustering, and then train a classifier on these labels. Our experiments reveal a strong positive correlation between the model's classification accuracy on a held-out ID test set and its OOD detection performance. Leveraging this correlation, we use the model's ID generalization ability as a direct criterion for hyperparameter selection, such as the choice of clustering method and the number of pseudo-labels. Importantly, unlike prior label-free OOD detection approaches that adapt powerful pre-trained feature extractors, our entire process is self-contained within the ID dataset and does not rely on knowledge transferred from external labeled corpora. With this automatic hyperparameter selection mechanism, our OOD detection method outperforms existing unsupervised baselines and is competitive with supervised approaches.

Furthermore, because pseudo-labels are not unique, where different numbers of clusters or different SSL backbones can induce distinct yet plausible classification logics, we introduce an ensemble strategy that aggregates these diverse logics to obtain a more robust and consistent final decision, further improving OOD performance. Overall, our findings challenge the notion that GT labels or knowledge transferred from other datasets are a necessary prerequisite for tackling challenging OOD tasks. Instead, we establish a strong empirical link between model generalization and OOD detection, thereby unifying supervised and unsupervised perspectives under a

This work is partially supported by the National Science and Technology Council (NSTC), Taiwan, under Contract: 113-2221-E-A49 -149 -MY3, 114-2221-E-A49 -129 -MY3, and 115-2425-H-A49 -001 -, and by the Higher Education Sprout Project of the National Yang Ming Chiao Tung University and Ministry of Education (MOE), Taiwan

common principle.

II. Related Works

Supervised OOD detection generally relies on learning discriminative decision boundaries from labeled ID data and then measuring how confidently test samples conform to this learned structure. Early work explored confidence-based scores [5], [6], distance-based metrics [7], and techniques for mitigating overconfidence [8], [9]. Subsequent studies improved robustness through contrastive feature learning and large-scale pretraining [10], [11], while post-hoc scoring strategies such as LogitNorm, MaxLogit, and MaxCosine [12], [13] further enhanced detection accuracy. More recent insights also highlight the effectiveness of feature-level norms derived from intermediate representations [14]. Together, these approaches reflect the broad trend of strengthening supervised OOD detection by refining confidence scoring and improving representation quality.

Unsupervised OOD detection removes the need for labeled data by modeling the structure of ID samples and identifying deviations from this structure. Reconstruction-based approaches using autoencoders or GANs [1], [15] and likelihood-based methods built on deep generative models [16] represent two primary lines of work. More recent methods leverage normalizing flows and transformer-based generative models [17], [18] to better capture complex data distributions, while diffusion-based reconstruction measures [4] further improve detection quality. Although these techniques reduce dependence on human annotation, they often incur higher computational cost and still trail supervised methods in overall performance.

Generating pseudo-labels through clustering has emerged as a promising direction for label-free OOD detection [19]–[21]. These methods typically use pseudo-labels to shape the learned representation and then apply post-hoc scoring rules, such as kNN distances or classifier confidence, to detect anomalies. Our approach differs in both objective and methodology. Unlike prior work, which often depends on pretrained features or focuses on designing a particular scoring function, our framework is fully self-contained and relies solely on the ID data. More importantly, we uncover a direct relationship between a model's generalization on the ID task and its OOD detection performance. This connection not only reframes the role of pseudo-labels as a means of inducing generalizable decision logic but also enables a practical advantage: ID generalization can be used as a principled signal for hyperparameter selection, such as choosing clustering methods or determining the number of pseudo-labels. This provides a simple yet powerful mechanism that existing pseudo-labeling methods do not exploit.

III. Methodology

Our central hypothesis is that a model's OOD detection ability emerges not from an independent mechanism but from how well it learns a consistent and generalizable decision logic on the ID data. In other words, stronger ID generalization should naturally yield more reliable OOD discrimination.

TABLE I: Correlation analysis between OOD detection performance and classification accuracy on the test set.

Method	FPR ↓	AUC ↑	Test Acc (%)
Random Assignment	46.56	84.25	10.72
Raw image clustering	31.87	93.22	94.17
GT labels	10.18	97.38	99.49

A. Pilot Study

To validate this hypothesis, we conducted a controlled pilot study using MNIST as the ID dataset and Fashion MNIST as the OOD dataset. We compared three supervision settings: random labels, clustering-based pseudo-labels, and ground-truth labels, and evaluated each model's ID generalization ability together with its OOD detection performance using FeatureNorm [14]. The goal of this comparison is not to optimize performance, but to directly test whether improvements in ID generalization translate into improvements in OOD detection. All implementation and evaluation details of this pilot study are provided in Appendix(Section I).

The results, summarized in Table I, reveal a clear monotonic trend: as the quality of the learned classification logic increased, so did the model's OOD detection performance. Models trained with random labels failed to separate ID and OOD samples, whereas those trained with ground-truth labels demonstrated both strong generalization and the highest OOD detection scores. These findings support our hypothesis and suggest that better pseudo-labeling strategies, particularly those based on richer representations such as SSL, may further strengthen both ID generalization and OOD performance.

B. Self-Supervised Pseudo-Label Generation

We utilize the mixed Barlow Twins method [22] to process distinct views of a batch of images X, denoted as Y^A and Y^B. These views are created by applying random augmentations $\mathcal{T}$ to X. As illustrated in Figure 1 (a), Y^A and Y^B pass through an encoder f_e and a projector f_p to generate normalized embeddings Z^A and Z^B. Additionally, a regularization branch processes interpolated images $Y^M = \lambda Y^A + (1-\lambda) Y^{B'}$, where $Y^{B'}$ represents a shuffled batch from Y^B. The output of this regularization branch is denoted as Z^M. These embeddings are then used to compute the Barlow Twins loss $\mathcal{L}_{BT}$ and Mixup [23] regularization objective $\mathcal{L}_{reg}$. The final loss function is:

$$\mathcal{L} = \mathcal{L}_{BT} + \lambda_{\text{reg}} \mathcal{L}_{\text{reg}}. \tag{1}$$

In the final loss, $\mathcal{L}_{BT}$ encourages similar representations for different augmentations of the same image, formulated as:

$$\begin{aligned} \mathcal{L}_{BT} = & \sum_i \left(1 - \frac{\left\langle z^A_{\cdot,i}, z^B_{\cdot,i} \right\rangle_b}{\left\| z^A_{\cdot,i} \right\|_2 \left\| z^B_{\cdot,i} \right\|_2} \right)^2 \\ & + \lambda_{BT} \sum_i \sum_{j \neq i} \left(\frac{\left\langle z^A_{\cdot,i}, z^B_{\cdot,j} \right\rangle_b}{\left\| z^A_{\cdot,i} \right\|_2 \left\| z^B_{\cdot,j} \right\|_2} \right)^2 \end{aligned} \tag{2}$$

where b indexes batch samples and i, j index feature dimensions. λ_{BT} is a weighting factor. The dot in $z^A_{\cdot,i}$ implies that all

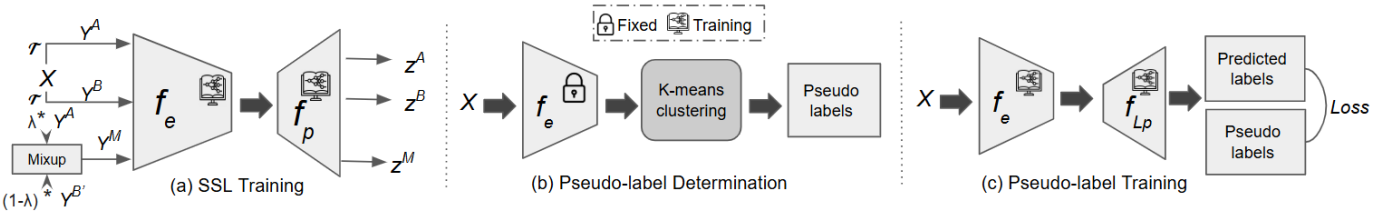

Fig. 1: An overview of our pipeline: (a) We first train a network using self-supervised learning to extract relevant features from the ID dataset. (b) Next, we apply K-means clustering to group the extracted features and generate pseudo-labels. (c) Finally, we train the network to classify images into pseudo-categories and compute feature norms to detect OOD samples.

features from the batch are considered. The first term of the Barlow Twins loss minimizes the variance of feature dimensions when aligning representations of different views of the same image, while the second term reduces redundancy between the dimensions.

The regularization term $\mathcal{L}_{reg}$ in final loss enhances Barlow Twins training by encouraging linearly interpolated inputs to produce linearly interpolated features. This adds regularization to the original objective. Let $Z^{M'} = \lambda Z^A + (1-\lambda)Z^{B'}$ represent features interpolated in the embedding space. The mixup regularization $\mathcal{L}_{reg}$ is defined as:

$$\begin{aligned}\mathcal{L}_{\text{reg}} = &\left\|(Z^M)^\top Z^A - (Z^{M'})^\top Z^A\right\|_2 \\ &+ \left\|(Z^M)^\top Z^B - (Z^{M'})^\top Z^B\right\|_2\end{aligned} \tag{3}$$

Here, Z^M refers to features derived from images interpolated in the image space, while $Z^{M'}$ refers to features interpolated in the embedding space. After SSL, the generated features are clustered using K-means, and the resulting clusters provide pseudo-labels for the classifier training.

C. Out-of-Distribution Data Detection based on Pseudo-labels

In this step, we trained a classifier (see Figure 1 (c)) to categorize images using the pseudo-labels generated from the clustering process. The underlying principle is similar to supervised OOD detection methods, where the model learns the internal structure of an ID dataset. Once the network is trained, we calculate the feature map norm of the convolution block [14] to classify samples as either ID or OOD. Specifically, given a feature map $z \in \mathbb{R}^{C\times W\times H}$ obtained from the trained classifier f_e, the average norm across all channels in a block B is computed as:

$$f_{\text{FN}}(x;B) = \frac{1}{C}\sum_{c=1}^{C}\sqrt{\sum_{w=1}^{W}\sum_{h=1}^{H}\max(z_c(w,h),0)^2}, \tag{4}$$

where $z_c \in \mathbb{R}^{1\times W\times H}$ represents the feature map of the c-th channel, and $\max(z_c(w,h),0)$ rectifies values by ignoring negative elements. This norm reflects the average activation level of the feature map for data x in block B. Following [14], samples with high norm values are considered ID, while those with low norm values are classified as OOD [24]. Additionally, we use the convolution block with the highest norm ratios, which is determined based on the response to ID data and their augmented variants, to distinguish ID from OOD samples.

D. Ensembling Pseudo-Labels

We explored the effect of ensembling in our method, as the pseudo-labels generated from different clustering processes can vary, while the GT labels remain unique. In this case, the ensemble networks differ not only in their initial seeds but also in the categories they learn, potentially offering diverse perspectives beyond networks trained with GT labels. Recall that $z \in \mathbb{R}^{C\times W\times H}$ represents a feature map. We compute the feature norm for channel c in model i as:

$$N_i^c = \sqrt{\sum_{w=1}^{W}\sum_{h=1}^{H}\max(z_c(w,h),0)^2}. \tag{5}$$

To ensemble the results, we aggregate the feature-norm vectors produced by multiple models. Specifically, let $N_i \in \mathbb{R}^{C\times 1\times 1}$ denote the feature-norm vector from the i-th model, where each entry N_i^c corresponds to the norm of channel c. We then compute the ensemble vector by taking the element-wise minimum across models:

$$N_{\text{min}} = \min_{i=1}^{k} N_i. \tag{6}$$

We then calculate the norm values $|N_{\text{min}}|$ to differentiate between ID and OOD samples.

IV. Experiments and Evaluations

We design experiments to validate our central hypothesis: a classifier can detect OOD samples because it has learned a consistent and generalizable classification logic. To test this principle, we train a classifier using pseudo-labels and select all hyperparameters based on the model's generalization performance on a held-out ID test set. We then assess the OOD detection performance of our approach against several baseline methods. For all experiments, every method is trained exclusively on the ID training data, with no external datasets utilized. After training, the models are evaluated on a combined test set of ID and OOD data. Furthermore, to ensure a fair comparison against baselines, our ensemble technique is not applied, except for the specific analysis in Section IV-E.

For evaluation,we report AUROC, a threshold-independent metric that measures performance across all possible thresholds, and FPR95 (False Positive Rate at 95% True Positive Rate), a

widely used metric that evaluates performance when the true positive rate is sufficiently high. All experiments are conducted on an NVIDIA RTX A4000 GPU. For inference, the processing time per batch of 16 images is 0.025 ± 0.007 seconds.

TABLE II: Comparison with unsupervised baselines under the OOD setup of [4]. Each column shows the average AUC over all OOD datasets for a given ID dataset. The best and second-best results are shown in **bold** and underlined.

ID	FMNIST	CIFAR-10	CelebA	SVHN	Avg. AUC
Likelihood [25]	38.4	39.7	43.8	50.2	43.0
WAIC [26]	38.4	39.7	43.7	50.2	42.9
DOS [27]	73.5	63.1	78.0	75.5	72.5
AutoEncoder	61.5	44.6	51.4	63.8	55.3
AE_MH [28]	79.1	44.3	61.1	77.7	65.6
MemAE [29]	54.9	43.5	54.5	74.0	56.7
AnoDDPM-Mod [30]	79.0	42.5	74.5	82.3	69.6
DDPM [4]	83.7	69.8	85.8	88.9	82.1
Our method	**84.7**	**79.0**	**86.2**	**90.8**	**85.2**

A. Evaluations on Benchmark Settings

We first assessed the performance of our pseudo-label strategy following the benchmark settings described in [4] in Table II. Specifically, two distinct configurations were employed: one for grayscale datasets and another for color datasets. In the grayscale setting, we used FashionMNIST as the ID dataset and MNIST as the OOD dataset. For the color setting, we evaluated three datasets: CIFAR-10, CelebA, and SVHN. Each dataset was treated as the ID dataset in turn, with the other two serving as OOD datasets, along with vertical and horizontal flipped versions of each ID dataset as extra OOD datasets. In comparison to other baseline methods, our approach demonstrated superior performance on average. Detailed results are provided in Table III-VIII of our supplementary material.

To achieve a realistic evaluation, we extended our analysis using ImageNet as the ID dataset and tested on four OOD datasets (i.e., INaturalist, SUN, Places, and Texture). For this experiment, we used ResNet-50 [31] as the backbone and employed the DINO SSL method [32], which achieved the highest k-NN accuracy on ImageNet. The average AUC and FPR values presented in Table III highlight the effectiveness of our approach. Notably, our approach outperformed NAN [24], despite NAN being an extension of FeatureNorm [14]. This is expected because NAN is tailored for specific conditions: (1) it is designed for use on the hidden layer of a projection head, and (2) it is optimized for classifiers trained with a contrastive loss. In contrast, we use more general setup, with a standard CNN backbone that ends in an average pooling layer and a classifier trained using conventional cross-entropy loss. Detailed results are provided in Section III-K of the supplementary material.

B. Choice of SSL Embeddings.

Since effective classifier learning relies on consistent decision logic across samples, the quality of pseudo-labels is critically dependent on the structure of the underlying embeddings. We therefore examine which SSL methods are most effective at producing high-quality embeddings. Among the SSL approaches, we compare several widely used ones, including SimCLR [33], BYOL [34], W-MSE [35], and mBT [22]. As shown in the upper part of Table IV, mBT achieves the best OOD detection performance. Noteably, this trend is consistent with the KNN test accuracy.

TABLE III: OOD detection performance on ImageNet. Averaged FPR and AUC values are reported.

Method	FPR↓	AUC↑
SSD [36]	71.77	77.16
KNN [37]	65.01	84.07
NAN [24]	51.08	87.61
Ours	**38.82**	**90.35**

To further validate whether selecting SSL embeddings based on KNN accuracy generalizes to large-scale datasets, we conduct additional experiments on ImageNet. Specifically, we use DINO to train three different backbones: ViT-S/8, ViT-S/16 [38], and ResNet-50 to obtain SSL embeddings and generate pseudo-labels. A ResNet-50 classifier is then trained on these pseudo-labels for OOD detection, using ImageNet as the ID dataset and keeping the OOD datasets unchanged. Results in the lower part of Table IV show that pseudo-labels from ViT-S/8 outperform those from ViT-S/16 and ResNet-50, confirming the generalizability of this criterion.

TABLE IV: Our experiment results suggest that KNN accuracy serves as a reliable indicator for selecting the appropriate SSL method and backbone.

ID Dataset	Backbone	SSL Method	FPR↓	AUC↑	KNN
CIFAR-10	ResNet18	SIMCLR	29.99	92.96	87%
		BYOL	27.15	93.67	88%
		W-MSE	27.45	93.18	88%
		mBT	**20.23**	**95.4**	**91%**
CIFAR-100	ResNet18	SIMCLR	67.29	79.38	57%
		BYOL	64.37	80.44	57%
		W-MSE	58.30	82.94	59%
		mBT	**54.92**	**86.21**	**62%**
ImageNet	ResNet-50	DINO	39.62	89.85	75%
	VIT(S/8)		**38.82**	**90.35**	**80%**
	VIT(S/16)		39.89	90.06	75%

C. Number of Pseudo-Categories.

We employed the K-means clustering method to generate pseudo-labels for training the classifier. The classifier's ability to detect OOD samples can depend on the number of pseudo-labels assigned to the ID dataset. To investigate the effect of the number of clusters, k, we conducted the experiment following the setup in [14]. The results, summarized in Table V, reveal that a smaller train-test accuracy gap indicates stronger generalization, highlighting the model's ability to capture meaningful classification logic. While the accuracy gap alone cannot pinpoint the optimal number of pseudo-categories, it serves as a valuable guideline for selecting a reasonable number of clusters to balance performance and generalization.

TABLE V: Impact of the number of pseudo-categories (K) on CIFAR-10/100. The ID train-test accuracy gap guides optimal cluster selection.

No.	CIFAR-10					CIFAR-100				
	Train	Test	Gap	FPR ↓	AUC ↑	Train	Test	Gap	FPR ↓	AUC ↑
K = 5	99.8	96.7	3.1	22.18	94.62	-	-	-	-	-
K = 10	98.5	95.1	3.4	24.46	94.01	94.0	88.7	5.3	68.23	74.50
K = 20	96.7	93.8	**2.9**	**20.23**	**95.40**	92.7	87.1	5.6	67.15	76.28
K = 50	94.0	89.3	4.7	43.28	90.08	90.5	85.4	5.1	61.78	79.26
K = 100	91.0	86.7	4.3	41.56	87.96	88.9	84.6	**4.3**	56.28	85.71
K = 200	89.1	80.7	8.4	60.42	79.02	83.5	79.0	4.5	**54.92**	**86.21**
K = 500	-	-	-	-	-	58.7	53.2	5.5	60.72	78.59

TABLE VI: Comparison of GT-labeled and pseudo-labeled ensembles. Parentheses indicate gains over the best individual model.

Metric	Best	Avg	Max	Min
	GT Labels			
AUC	93.79	94.01 (+0.22)	93.18 (-0.61)	94.29 (+0.50)
FPR	26.32	26.08 (-0.24)	27.13 (+0.81)	25.85 (-0.47)
	Pseudo Labels			
AUC	95.4	95.83 (+0.43)	94.68 (-0.72)	**96.35** (+0.95)
FPR	20.23	18.84 (-1.39)	21.83 (+1.6)	**17.85** (-2.38)

D. Pseudo-Labels visualization

To provide further insight into how the model organizes visual information when trained with pseudo-labels, we present a t-SNE visualization in Figure 2. The visualization shows that pseudo-labels and ground-truth (GT) labels share substantial overlap at a coarse level, while exhibiting meaningful differences in how fine-grained visual structure is organized. In particular, images of birds (green dots) are separated into two pseudo-categories, corresponding to full-body views and head-only views, respectively. A similar pattern is observed for horses (gray dots), where one pseudo-category captures side-view full-body images, while another focuses on frontal or partial-body appearances. This observation supports our broader argument that pseudo-labels provide a consistent classification logic that is sufficient for learning a meaningful decision boundary. Please refer to Figure 3 in the supplementary material for additional details and a high-resolution version.

E. Effectiveness of Ensembling Pseudo-Labels

To study the effect of ensembling with pseudo-labels, we trained three models on CIFAR-10 using 5, 10, and 20 pseudo-categories, chosen for their comparable detection performance (Table V). For comparison, we also trained three GT models with different random seeds. Unlike GT labels, which are fixed and provide no diversity in learning perspectives, pseudo-labels derived from different SSL configurations or varying numbers of pseudo-categories induce distinct classification logics. Ensembling across such diverse decision boundaries can therefore improve OOD detection.

In addition to the N_{min} ensemble introduced in Section III-D, we further evaluate two variants: N_{max} and N_{avg}, obtained by replacing the minimum operation with maximum and average, respectively. The resulting norm values are then used to distinguish ID from OOD samples.

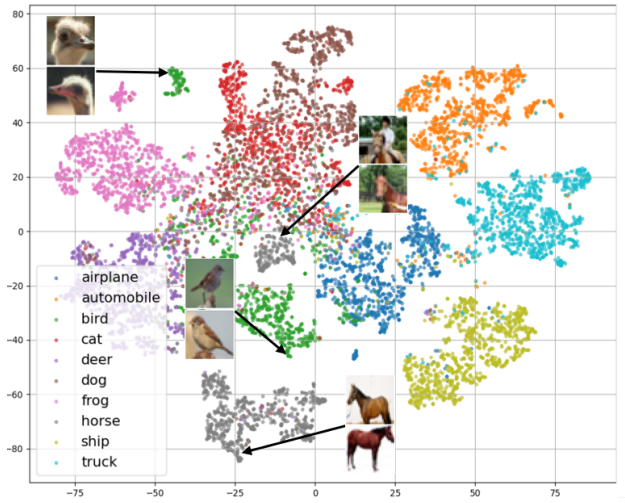

Fig. 2: t-SNE [39] visualization of CIFAR-10 embeddings learned with pseudo-labels.

As shown in Table VI, ensembling with pseudo-labels outperformed ensembling with GT labels. Among all the strategies, the minimum ensemble delivered the best performance. This result is intuitive: models tend to produce larger feature norm values for ID samples due to the familiarity of the data's structure [24], whereas OOD samples generate smaller feature norm values. By focusing on the minimum value ensemble, the smaller norms of OOD samples are emphasized, making them more distinguishable from the larger norms of ID data. This strategy reduces false positives and improves OOD detection performance compared to other ensemble approaches.

F. Extended Experimental Evaluation

Additional experiments are provided in the supplementary material, including comparisons on adjacent OOD datasets (Section III-A), the rationale for using FeatureNorm as the OOD scoring method (Section III-B), analysis of semantic and fine-grained structures in pseudo-label embeddings (Section III-C), the necessity of SSL encoder fine-tuning (Section III-D), ensembling with pseudo-labels on ImageNet (Section III-E), OOD detection for flipped images (Section III-F), quality of pseudo labels (Section III-G), choice of clustering methods (Section III-H), model generalization (Section III-I) and detailed results for individual OOD datasets (Sections III-J and III-K). These analyses further validate our approach and demonstrate its versatility across diverse scenarios.

G. Limitations

Our work's central insight is that a model's OOD detection capability emerges from its ability to learn a consistent and generalizable classification logic. While we primarily validate this hypothesis through empirical experiments, a more formal theoretical foundation remains an open direction. We also observe that classification logic derived from GT labels tends

to outperform that obtained from pseudo-labels, particularly in challenging adjacent-OOD scenarios. Nevertheless, this gap indicates a valuable direction for future research to develop more reliable pseudo-labeling strategies that maintain class-level semantic consistency even near decision boundaries.

V. Conclusions

This work reexamines OOD detection by showing that a model's ability to detect outliers is fundamentally driven by its generalization on the ID task. Rather than depending on GT labels, effective OOD detection arises from learning a consistent and generalizable classification logic. To validate this idea, we introduced a framework that generates pseudo-labels through SSL and clustering. Experiments confirmed a strong positive link between ID generalization and OOD detection, and our ensemble method further improved performance. Our findings suggest that, alongside efforts to design new scoring functions, enhancing models' generalization on ID data offers a promising path for advancing unsupervised OOD detection.

References

[1] Jinwon An and Sungzoon Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Special lecture on IE*, vol. 2, no. 1, pp. 1–18, 2015.

[2] Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios, "Neural spline flows," *NeurIPS*, vol. 32, 2019.

[3] Hamidreza Kamkari, Brendan Leigh Ross, Jesse C Cresswell, Anthony L Caterini, Rahul G Krishnan, and Gabriel Loaiza-Ganem, "A geometric explanation of the likelihood ood detection paradox," *arXiv preprint:2403.18910*, 2024.

[4] Mark S Graham, Walter HL Pinaya, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin, and Jorge Cardoso, "Denoising diffusion models for out-of-distribution detection," in *CVPR*, 2023, pp. 2947–2956.

[5] Dan Hendrycks and Kevin Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint:1610.02136*, 2016.

[6] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," *arXiv preprint:1706.02690*, 2017.

[7] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," *NeurIPS*, vol. 31, 2018.

[8] Terrance DeVries and Graham W Taylor, "Learning confidence for out-of-distribution detection in neural networks," *arXiv preprint:1802.04865*, 2018.

[9] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li, "Energy-based out-of-distribution detection," *NeurIPS*, vol. 33, pp. 21464–21475, 2020.

[10] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin, "Csi: Novelty detection via contrastive learning on distributionally shifted instances," *NeurIPS*, vol. 33, pp. 11839–11852, 2020.

[11] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan, "Exploring the limits of out-of-distribution detection," *NeurIPS*, vol. 34, pp. 7068–7081, 2021.

[12] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li, "Mitigating neural network overconfidence with logit normalization," in *ICML*, 2022, pp. 23631–23644.

[13] Zihan Zhang and Xiang Xiang, "Decoupling maxlogit for out-of-distribution detection," in *CVPR*, 2023, pp. 3388–3397.

[14] Yeonguk Yu, Sungho Shin, Seongju Lee, Changhyun Jun, and Kyoobin Lee, "Block selection method for using feature norm in out-of-distribution detection," in *CVPR*, 2023, pp. 15701–15711.

[15] Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *IPMI*, 2017, pp. 146–157.

[16] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan, "Likelihood ratios for out-of-distribution detection," *NeurIPS*, vol. 32, 2019.

[17] Evan D Cook, Marc-Antoine Lavoie, and Steven L Waslander, "Feature density estimation for out-of-distribution detection via normalizing flows," *arXiv preprint:2402.06537*, 2024.

[18] Alexander Podolskiy, Dmitry Lipin, Andrey Bout, Ekaterina Artemova, and Irina Piontkovskaya, "Revisiting mahalanobis distance for transformer-based out-of-domain detection," in *AAAI conference on artificial intelligence*, 2021, vol. 35, pp. 13675–13682.

[19] Niv Cohen, Ron Abutbul, and Yedid Hoshen, "Out-of-distribution detection without class labels," in *ECCV*, 2022, pp. 101–117.

[20] Di Jin, Shuyang Gao, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tür, "Towards textual out-of-domain detection without in-domain labels," *TASLP*, vol. 30, pp. 1386–1395, 2022.

[21] Byounghan Lee, Jaesik Kim, Junekyu Park, and Kyung-Ah Sohn, "Improving unsupervised out-of-domain detection through pseudo labeling and learning," in *Findings of the Association for Computational Linguistics: EACL 2023*, 2023, pp. 1031–1041.

[22] Wele Gedara Chaminda Bandara, Celso M De Melo, and Vishal M Patel, "Guarding barlow twins against overfitting with mixed samples," *arXiv e-prints*, pp. arXiv–2312, 2023.

[23] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint:1710.09412*, 2017.

[24] Jaewoo Park, Jacky Chen Long Chai, Jaeho Yoon, and Andrew Beng Jin Teoh, "Understanding the feature norm for out-of-distribution detection," in *ICCV*, 2023, pp. 1557–1567.

[25] Christopher M Bishop, "Novelty detection and neural network validation," in *International Conference on Artificial Neural Networks*. Springer, 1993, pp. 789–794.

[26] Hyunsun Choi, Eric Jang, and Alexander A Alemi, "Waic, but why? generative ensembles for robust anomaly detection," *arXiv preprint:1810.01392*, 2018.

[27] Warren Morningstar, Cusuh Ham, Andrew Gallagher, Balaji Lakshminarayanan, Alex Alemi, and Joshua Dillon, "Density of states estimation for out of distribution detection," in *AISTATS*, 2021, pp. 3232–3240.

[28] Taylor Denouden, Rick Salay, Krzysztof Czarnecki, Vahdat Abdelzad, Buu Phan, and Sachin Vernekar, "Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance," *arXiv preprint:1812.02765*, 2018.

[29] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *ICCV*, 2019, pp. 1705–1714.

[30] Julian Wyatt, Adam Leach, Sebastian M Schmon, and Chris G Willcocks, "Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise," in *CVPRW*, 2022, pp. 650–656.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[32] 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, pp. 9650–9660.

[33] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020, pp. 1597–1607.

[34] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al., "Bootstrap your own latent-a new approach to self-supervised learning," *NeurIPS*, vol. 33, pp. 21271–21284, 2020.

[35] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe, "Whitening for self-supervised representation learning," in *ICML*, 2021, pp. 3015–3024.

[36] Vikash Sehwag, Mung Chiang, and Prateek Mittal, "Ssd: A unified framework for self-supervised outlier detection," *arXiv preprint:2103.12051*, 2021.

[37] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li, "Out-of-distribution detection with deep nearest neighbors," in *ICML*, 2022, pp. 20827–20840.

[38] 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:2010.11929*, 2020.

[39] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *JMLR*, vol. 9, no. Nov, pp. 2579–2605, 2008.