Sharpness-aware Fine-Tuning for OOD Detection

Chi Zhang, Wei Wang, *Member, IEEE,* Yao Zhao, *Fellow, IEEE,* Nicu Sebe, *Senior Member, IEEE,* Yue Song, *Member, IEEE*

***Abstract*—The out-of-distribution (OOD) detection task is crucial for the real-world deployment of machine learning models. In this paper, we propose to study the problem from the perspective of Sharpness-aware Minimization (SAM). Compared with traditional optimizers such as SGD, SAM can better improve the model performance and generalization ability, and this is closely related to OOD detection [1]. Therefore, instead of using SGD, we propose to fine-tune the model with SAM, and observe that the score distributions of in-distribution (ID) data and OOD data are pushed away from each other. Besides, with our carefully designed loss, the fine-tuning process is very time-efficient. The OOD performance improvement can be observed after fine-tuning the model within 1 epoch. Moreover, our method is very flexible and can be used to improve the performance of different OOD detection methods. Extensive experiments have demonstrated that our method achieves *state-of-the-art* performance on widely-used OOD benchmarks across different architectures. Moreover, comprehensive ablation studies and theoretical analyses are discussed to support the empirical results.**

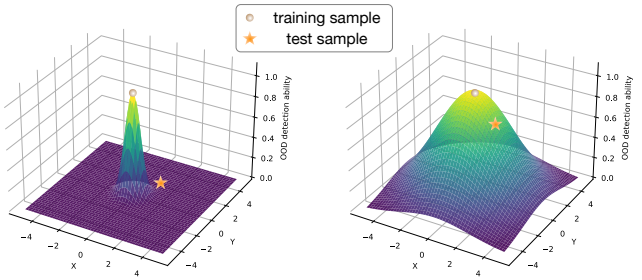
***Index Terms*—Out-of-distribution Detection, Sharpness-aware Minimization**

I. Introduction

DEEP learning models are known to be fragile to distribution shifts. Detecting Out-of-Distribution (OOD) data is thus important for the real-world deployment of machine learning models [2]. The major difficulty arises from the fact that existing deep learning models can make excessively confident but incorrect predictions for OOD data [3].

To address this challenge, previous *post hoc* methods distinguish in-distribution (ID) and OOD data either by using feature distance [4], gradient abnormality [5], or logit manipulation [6]. Although these *post hoc* methods do not need to fine-tune or re-train the models, the performances are usually inferior compared with the supervised ones which require further training/fine-tuning. Also, *post hoc* methods might be confined to certain network architectures and benchmarks.

From the perspective of the availability of extra training/fine-tuning data, the supervised methods can be divided into three types: (1) no extra data; (2) using real OOD data; (3) using pseudo-OOD data (generated from ID data). For the first type of methods, Huang and Li [7] proposed MOS to fine-tune the pretrained model. By reducing the number of in-distribution classes through grouping similar semantic categories, the model can simplify the decision boundary and reduce the uncertainty space between ID and OOD data. However, this method requires additional human labor to group similar semantic categories and select the suitable number of groups. The KL Matching [8] method also belongs to the first type. It learns a new distribution for each class from the validation ID dataset to enhance multi-class OOD detection. However, these methods usually have inferior performance compared with the other two types of methods which use extra OOD data.



(a) SGD (b) SAM

Property	SFT	FT(SGD)
sharpness	2.0591	2.1255
norm of $\nabla f(\boldsymbol{w})$ change	1.4284	1.4761

(c) Validation of different optimizers on ResNetv2-101.

Fig. 1. The motivation of our SFT. Subfigures (a) and (b) show the variation in OOD detection ability of SGD and SAM when the test OOD distribution shifts away from the training OOD distribution. For visualization purposes, the data are reduced to two dimensions on the x and y axes, with the z-axis capturing OOD detection performance (inverse of OOD loss). And the metrics for sharpness and gradient norm change in subfigure (c) confirm that SAM converges to the smoother loss landscape, as lower values in both metrics denote greater smoothness.

For instance, as a representative of the second type of methods, OE [9] requires additional training with extra real OOD data. However, it has quite a few inevitable drawbacks, such as its reliance on a large amount of real OOD data (*i.e.* real OOD images rather than pseudo ones), costly training time which usually takes plenty of epochs (>5), and poor robustness across different architectures and benchmarks. These drawbacks limit the practical application of these methods which use real OOD data.

To avoid using real OOD data, the third type of methods has been proposed which uses pseudo-OOD data. These methods are more user-friendly and have practical applications. Usually, they have better performance than the first type. Taking the FeatureNorm [10] method for example, it uses the pseudo-OOD data to select the block with best OOD detection ability, then computes the feature norm of the selected block as OOD detection score. However, its performance is slightly inferior

Chi Zhang, Wei Wang, Yao Zhao are with the Institute of Information Science, Beijing Jiaotong University, Beijing, China. E-mail: chi.zhang@bjtu.edu.cn, wei.wang@bjtu.edu.cn, yzhao@bjtu.edu.cn.

Nicu Sebe is with the Department of Information Engineering and Computer Science (DISI), University of Trento, Italy. E-mail: sebe@disi.unitn.it

Yue Song is with the College of AI, Tsinghua University, China. E-mail: songyue19960927@gmail.com

Manuscript received April 19, 2021; revised August 16, 2021.

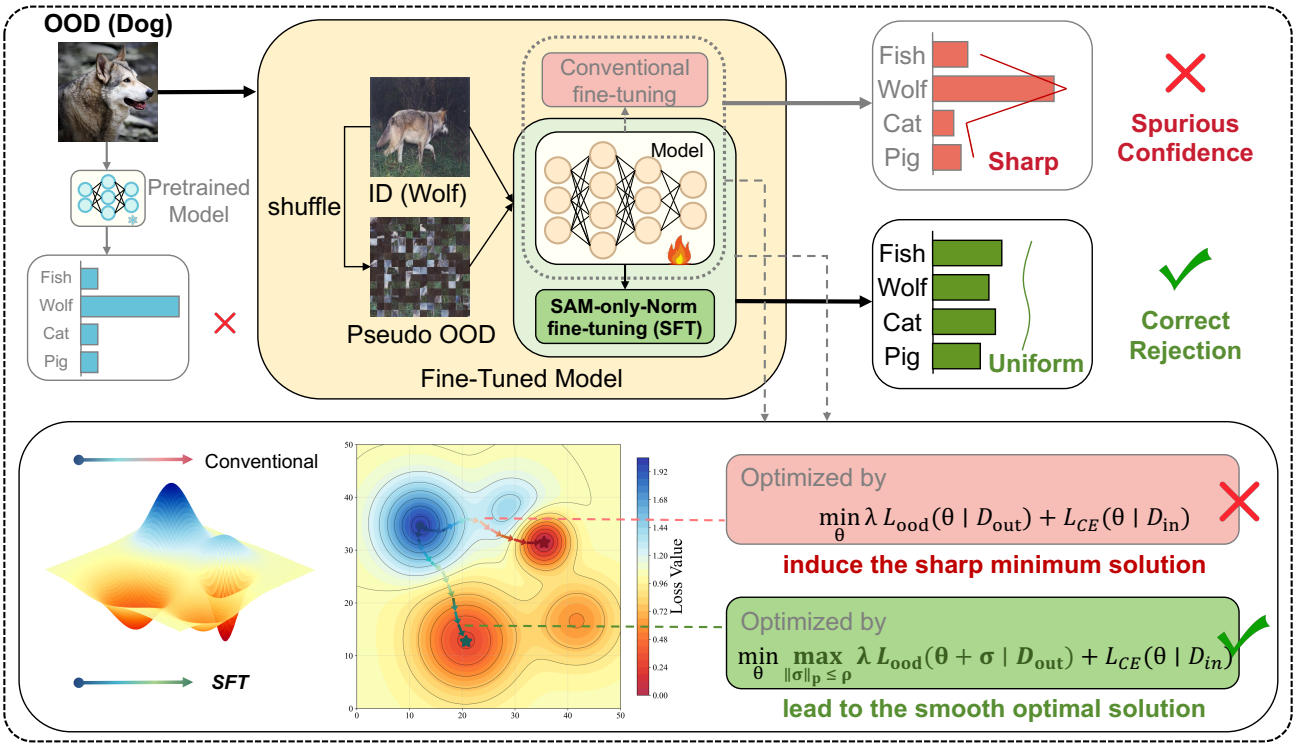


Fig. 2. Our SFT fine-tunes the pretrained model using ID and pseudo-OOD data. In the fine-tuning process, SFT could guide the model to achieve the smoother loss landscape by minimizing the SAM loss term, which naturally enables robust OOD detection. As demonstrated by the example at the top, the smoother loss landscape finally leads to better score distributions for OOD and eventually helps to better distinguish the ID and OOD samples.

compared with the second type of methods which uses real OOD data. Due to its practical application, this paper focuses on the third type of methods which fine-tune the pretrained model with pseudo-OOD data.

The majority of third-type approaches apply standard optimizers such as SGD or Adam to regularize the pretrained model. For instance, DreamOOD [11] leverages the diffusion model to generate pixel-level pseudo-OOD data and fine-tunes the original model with SGD. Despite their widespread use, the above methods often overlook the critical role of the optimizer during the fine-tuning process. Especially when pseudo-OOD data supervision is involved, different optimization strategies may drastically affect the OOD detection performance of the fine-tuned model. Hence, investigating and exploring optimizer selection is essential for OOD detection task. Recently, Sharpness-aware Minimization (SAM) [12] is one well-known smoothness optimization technique which indicates a novel training framework by simultaneously minimizing the loss value and the sharpness. The concept of sharpness defines the smoothness of the data point in the neighborhood of the loss landscape. *Intuitively, the model fine-tuned with SAM can better handle diverse OOD samples during testing, as SAM enforces loss smoothness within a hypersphere around each model parameter during OOD fine-tuning process. This enhances robustness and enables the model to assign lower confidence for diverse and hard OOD inputs.* The intuition outlined above is illustrated in Fig. 1: for training OOD samples (depicted as spheres), both SGD and SAM can achieve nearly perfect OOD detection performance after careful OOD fine-tuning. The key distinction between different optimizers lies in the smoothness of the loss landscape of fine-tuned model shown in Fig. 1. However, in practical OOD detection scenarios, test OOD samples (shown as stars) typically deviate slightly from the training OOD data. As shown in Fig. 1, the smoother loss surface induced by SAM allows it to better withstand such deviations, thereby yielding better performance across diverse OOD test datasets compared to conventional optimizers. We argue this clue could be leveraged to further enhance the OOD detection ability of model.

Motivated by the above, we propose our Sharpness-aware Fine-Tuning (SFT) method, in which the pretrained model is fine-tuned by SAM using pseudo-OOD data within 1 epoch (50 steps more precisely). Fig. 2 depicts the overall procedure of the fine-tuning process. Our SFT generates pseudo-OOD data using both *Jigsaw Puzzle Patch Shuffling* and *RGB Channel Shuffling*. As can be seen in Fig. 3, ID data and OOD data are better distinguished. Our SFT has improved the OOD detection performance of typical baseline OOD methods (*i.e.,* Energy [6], MSP [13], ASH [14] and RankFeat [2]) by a large margin, boosting their performance 14.07% in FPR95 and 4.04% in AUROC on average. In particular, for the latest *state-of-the-art* method VRA [15], our method can greatly improve the performance to gain new *state-of-the-art* performance with 27.41% in FPR95 and 94.36% in AUROC on average. Moreover, for some *post-hoc* methods which are originally sensitive to specific architectures, our fine-tuning method can alleviate such a problem and improve their performance to a competitive level. For instance, GradNorm [5] which originally has a 92.68% FPR95 on Textures can achieve a 51.03% FPR95 after fine-tuning. Extensive ablation studies are performed to reveal important insights of SFT, and concrete theoretical analysis is

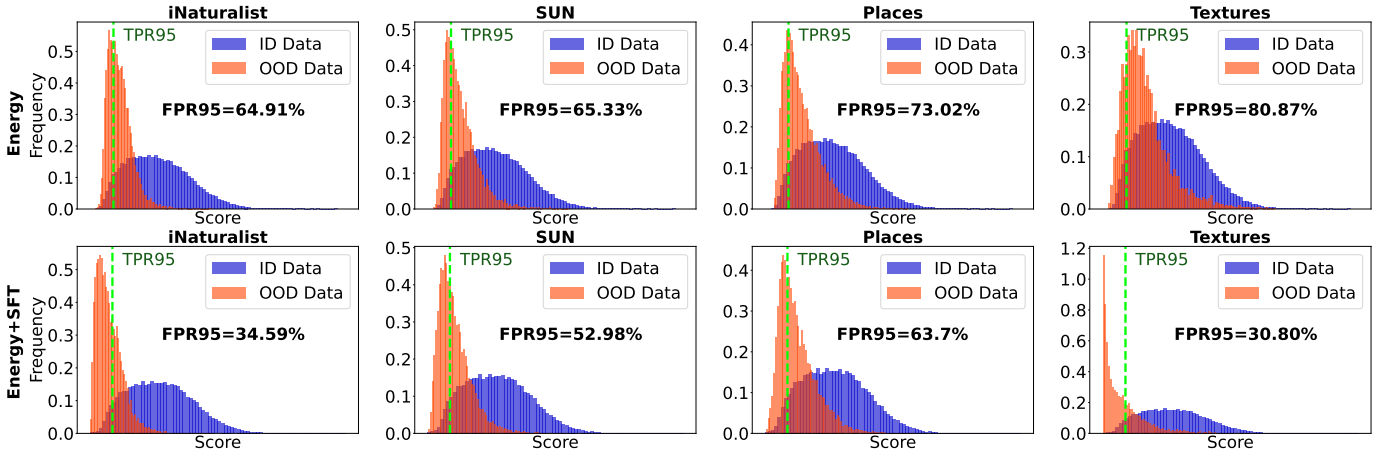


Fig. 3. The score distributions of `Energy` [6] (*top row*) and our proposed `SFT+Energy` (*bottom row*) on four OOD datasets. The dashed green line indicates the discrimination threshold employed for OOD evaluation, which is set at TPR95. As can be seen, ID and OOD data could be separated better utilizing `SFT` under the shown discrimination threshold, which leads to lower FPR95. Moreover, our `SFT` has improved the OOD detection performance of Energy with 30.32%, 12.35%, 9.32%, and 50.07% in FPR95 on iNaturalist, SUN, Places, and Textures. The consistent performance across various OOD datasets demonstrates the robustness of our `SFT`.

conducted to explain the underlying mechanism and support our empirical observation.

The key results and main contributions are threefold:

- We propose Sharpness-aware Fine-Tuning (`SFT`), a simple, real OOD data-free, and time-efficient method for OOD detection by fine-tuning pre-trained model with SAM using generated pseudo-OOD data within 1 epoch. Comprehensive experimental results indicate that our `SFT` brings significant improvements in various OOD detection benchmarks and methods on the conventional visual models and recently widespread CLIP.
- We perform extensive ablation studies on the impact of (1) the advantage of SAM over other optimizers; (2) the efficacy of pseudo-OOD data; (3) comparison with real OOD data based methods; (4) the effectiveness of SAM-only norm; (5) the perturbation radius ρ in SAM. These results motivate how we design our method and tune the hyper-parameters.
- Theoretical analyses are conducted to shed light on the working mechanism: from the perspective of Robust Machine Unlearning, we formulate the OOD fine-tuning process as a min-max robust optimization problem. This formulation naturally connects SAM framework with robust OOD detection, thereby explaining why SAM optimizer is uniquely effective for OOD detection. Moreover, we offer the theoretical explanation for the improved OOD performance by visualizing the two vital terms in the upper bound of the `Energy` score change.

The above observations also hold for other architectures on different testsets, which validates our theoretical analysis. Please refer to Supplementary Material F for details.

II. Related Work

A. Sharpness-aware Minimization.

Optimizers play important roles in multiple machine learning tasks, such as SGD and Adam. However, these traditional optimizers are unable to search for global optima that is smooth enough on the loss landscape. To overcome this drawback, Sharpness-aware Minimization (SAM) indicates a novel training framework by simultaneously minimizing the loss value and the sharpness [12]. Over the years, SAM has been proven to be efficient in various settings [16], [17]. There are several adaptations of the original SAM that further enhance the optimization, such as ASAM [18], GSAM [19], SAM-only-norm [20], and SSAM [21]. Another alternative research direction addresses the issue of rising computational expenses and proposes different acceleration techniques, including LookSAM [22] and ESAM [23]. Of equal significance, a substantial amount of research work strives to elucidate the working mechanism of SAM from different perspectives, such as [24], and [25]. Motivated by the above research findings, our `SFT` enhances the various *post-hoc* methods by smoothing the loss landscape of the model.

B. Machine Unlearning.

Machine unlearning aims to eliminate the influence of unwanted data by modifying the model, and they were initially developed to reduce post-training privacy risks [26]. Although retraining the model from scratch can guarantee exact unlearning, this method is often infeasible due to its high computational cost. Consequently, the research focus has shifted towards approximate unlearning methods [27], [28]. The core idea of these methods is to strike a balance between efficiency and effectiveness by directly modifying the pre-trained model to approximate the effect of the full retraining.

In recent years, with the development of Large Language Models (LLMs), the need to perform unlearning on them has become increasingly strong for the sake of Trustworthy AI and AI Safety [29], [30]. Existing LLM unlearning methods can be broadly divided into two categories: (1) Model-based optimization: These methods achieve unlearning by directly modifying the parameters of model [31], [32]; (2) Input-based approaches: These approaches leverage techniques such as prompt and in-context learning to attain unlearning at inference time, avoiding substantial parameter updates [33],

[34]. Nevertheless, recent studies have revealed that existing unlearning methods struggle to achieve robust unlearning [35], [36]. Several studies have investigated robust unlearning. For instance, [30] integrates the SAM optimizer to make the unlearning process more robust from the optimization standpoint. In light of the above research findings, we reconsider the OOD fine-tuning as the specific machine unlearning process. This formulation naturally connects SAM framework with robust OOD detection, thereby explaining why SAM optimizer is uniquely effective for OOD detection.

C. Distribution shift.

The consistent presence of distribution shifts has posed an ongoing challenge in the field of machine learning research [37]–[43]. The issue of distribution shifts can be broadly categorized into two types: shifts in the input space and shifts in the label space. When distribution shifts occur specifically in the input space, they are commonly referred to as covariate shifts [44], [45]. In this setting, the input data is subject to perturbations or domain shifts, while the label space remains unchanged [46]. The primary objective is to enhance the robustness and generalization of the model [47]. When it comes to the task of OOD detection, the labels are mutually exclusive. The primary goal is to ascertain whether a test sample should be predicted by the pre-trained model [48]. From the perspective of distribution shift, we generate the pseudo-OOD data to mimic the distribution of OOD samples. And we conduct theoretical analyses to shed light on the working mechanism of our SFT from the standpoint of distributional shift in the Supplementary Material.

D. OOD Detection.

There are three streams of OOD detection approach: generative model based, discriminative model based and large language model based. Generative model usually leverage the likelihood of the data for distinguishing OOD samples [49], [50]. Here we mainly highlight discriminative model based OOD detection and large language model based OOD detection. For discriminative model based OOD detection, one major direction of OOD detection is to design effective scoring functions, mainly consisting of confidence-based methods [13], [51], [52], distance-based approaches [4], [53]–[55], energy-based score [2], [6], [56]–[58], gradient-based score [5], and Bayesian methods [59]. Another promising line of OOD detection is to regularize the pre-trained model to produce lower confidence on OOD data [9], [60]–[62]. Compared with *post hoc* scoring functions, these approaches usually achieve superior OOD detection performance. However, the above methods often require extra real OOD data and long training time, which limits their deployment in real-world applications. In contrast, our SFT is both OOD-data-free and time-efficient, which can be successfully applied to various OOD detection baselines.

Recently, leveraging large language models (LLMs) for OOD detection is also an emerging trend [63]–[67]. Among these methods, zero-shot OOD detection methods assisted by CLIP [68] have achieved impressive OOD detection performance, such as MCM [63] and ZOC [64]. By carefully crafting prompts, prompt learning methods can boost the few-shot OOD detection performance of CLIP, including CoOp [69], LoCoOp [65], NegPrompt [67] and GalLoP [70]. [71] approaches few-shot OOD detection from both textual and visual perspectives, further enhancing it by integrating the prior knowledge of CLIP. However, the above CLIP-based OOD detection methods have inferior performance due to the absence of information from OOD data. By utilizing the generated pseudo-OOD data and sharpness-aware optimization, our SFT can mimic the real OOD data and optimize the loss to a flat landscape to improve the few-shot OOD detection.

III. Methodology

In this section, we first introduce the OOD detection task and Sharpness-aware Minimization, then present our proposed SFT that performs OOD detection by fine-tuning with SAM using pseudo-OOD data.

A. Preliminary & Background knowledge

1) Preliminary: OOD detection task.: The OOD detection problem usually is defined as a binary classification problem between ID data $\mathcal{D}_{id}$ and OOD data $\mathcal{D}_{ood}$. For a given model F which is trained on ID data, OOD detection aim to design the score function $\mathcal{G}$:

$$\mathcal{G}(\mathbf{x}) = \begin{cases} \text{in} & \mathcal{S}(\mathbf{x}) > \gamma, \\ \text{out} & \mathcal{S}(\mathbf{x}) < \gamma. \end{cases} \tag{1}$$

where x denotes the sample to be tested, $\mathcal{S}(.)$ is the seeking score function. γ is the chosen threshold to distinguish whether the given sample is ID or OOD. The most crucial challenge in OOD detection is how to design an effective score function to distinguish between the two types of data.

2) Background knowledge: Sharpness-aware Minimization.: SAM strives to minimize the loss value and smooth the loss landscape [12]. For the original given loss $L(\boldsymbol{w})$, the following is the corresponding Sharpness-Aware Minimization (SAM) objective of optimization:

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & L_{\mathcal{S}}^{SAM}(\boldsymbol{w}) + \lambda ||\boldsymbol{w}||_2^2 \quad , \\ where \quad & L_{\mathcal{S}}^{SAM}(\boldsymbol{w}) \triangleq \max_{||\boldsymbol{\epsilon}||_p \leq \rho} L_S(\boldsymbol{w} + \boldsymbol{\epsilon}). \end{aligned} \tag{2}$$

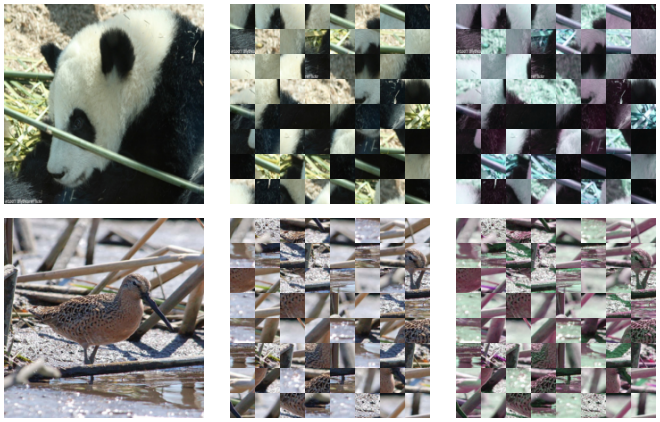
where ρ is the perturbation radius and ϵ can be computed by the following equation:

$$\hat{\epsilon} = \rho \operatorname{sign}\left(\nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w})\right) |\nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w})|^{\boldsymbol{q}-1} / \left(||\nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w})||_q^q\right)^{1/p}. \tag{3}$$

We set $p = q = 2$ following the setting of Foret et al. [12]. Then we can calculate the derivative of the model parameters $\boldsymbol{w}$ to obtain our final gradient approximation:

$$\nabla_{\boldsymbol{w}} L_{\mathcal{S}}^{SAM}(\boldsymbol{w}) \approx \nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w})|_{\boldsymbol{w}+\hat{\boldsymbol{\epsilon}}(\boldsymbol{w})}. \tag{4}$$

After obtaining the gradient, we can use a certain base optimizer to complete the entire model optimization process, such as SGD and Adam.



(a) (b) (c)

Fig. 4. Visualization of pseudo OOD data on ImageNet-1K [72]. (a) Original ID data; (b) *Jigsaw Puzzle Patch Shuffling*; and (c) *Jigsaw Puzzle Patch Shuffling+RGB Channel Shuffling*

B. SFT: Sharpness-aware Fine-Tuning

1) Generation of the pseudo-OOD data.: To avoid relying on real OOD data, we generate the pseudo-OOD data by *Jigsaw Puzzle Patch Shuffling* and *RGB Channel Shuffling*. Some examples are visualized in the Fig. 4. Here we provide a detailed introduction of these two methods.

- *Jigsaw Puzzle Patch Shuffling*. For the original ID images of dimensions $H \times W$, we first divide the original ID image into small patches. Then, we randomly shuffle and reassemble these small patches to generate pseudo-OOD data of the same dimensions $H \times W$. To introduce some randomness, we further randomly choose the used patch size at each fine-tuning step.
- *RGB Channel Shuffling*. To generate spatially coherent pseudo-OOD data, we randomly shuffle the RGB channels of the image at each fine-tuning step.

Our pseudo-OOD data generation aims to disrupt the high-level semantics of the original ID data in spatial structure and color appearance while preserving its original texture features. Specifically, *Jigsaw Puzzle Patch Shuffling* is responsible for destroying the overall shape and spatial relationships of objects, while *RGB Channel Shuffling* eliminates their original color patterns. Pseudo-OOD data generated by simultaneously disrupting spatial structure and color patterns are located far from the original ID data in the feature space. During the OOD fine-tuning process, this provides the clear signal that enables the model to learn distinct OOD detection decision boundary. The ablation study in Table I of the Supplementary Material further validate that the combination of the two methods for pseudo-data generation is more beneficial for OOD fine-tuning.

2) Sharpness-aware Fine-Tuning (SFT).: The SAM optimizer is proposed to minimize the loss value and the loss sharpness in the optimization landspace. Recently, it has been proven to be applicable to multiple machine learning tasks [12]. Motivated by OE [9], we fine-tune the pre-trained model using SAM to push the ID and OOD score distributions away. Similar to OE, given the data distributions $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$, we fine-tune the pre-trained model f using the loss function $L_{\texttt{SFT}}$ defined as:

$$L_{\texttt{SFT}}=\mathbb{E}_{(x,y)\sim\mathcal{D}_{in}}L_{CE}(f(x),y)+\lambda\mathbb{E}_{x'\sim\mathcal{D}_{out}}D(f(x'),U). \quad (5)$$

where $L_{CE}(f(x),y)$ is the cross-entropy loss. The key to improving the OOD detection performance is to minimize $D(f(x'),U)$ which measures the statistical distance between the output of OOD samples and the uniform distribution U. We use KL divergence to evaluate this distance in our experiments. To craft $\mathcal{D}_{out}$, we use both *Jigsaw Puzzle Patch Shuffling* and *RGB Channel Shuffling* methods to generate pseudo-OOD data from the orginal ID data.

The original SAM enforces perturbations to all the parameters of the model. This may introduce strong interference and negatively influence the OOD detection performance. We give a preliminary analysis of this phenomenon in the ablation study of Sec. V-C. Therefore, we choose SAM-only-norm [20], an improved version of SAM which only perturbs the normalization layers, as our optimizer throughout the experiments and theoretical analysis.

IV. Theoretical Analysis

In this section, we provide the theoretical analysis for the efficacy of our SFT from two distinct perspectives. First, we explain why our SFT could improve the OOD detection performance of the model from the perspective of the upper bound. Second, we offer the theoretical analysis from the standpoint of robust machine unlearning to explain why our SFT is particularly suitable for OOD fine-tuning. Together, these two distinct perspectives provide the solid theoretical guarantee for our SFT.

A. Upper Bound Analysis

In this part, we first present the upper bound of the change of Energy score at each fine-tuning iteration. Then we illustrate the changes of the upper bound formula by analyzing the specific terms through numerical experiments. From the perspective of the upper bound, we explain why SFT could improve the OOD performance of the model.

Proposition 1: The upper bound of Energy score's change during the SAM-only-norm fine-tuning process is defined as:

$$\triangle(\texttt{Energy}(x)) \leq -\frac{h}{2}\mathbb{E}\|\nabla f(w^t)\|^2+\triangle(max(g_i))+S \quad (6)$$

where $\triangle(\texttt{Energy}(x)) = log\sum_{i=1}^{n}e^{g_i^{t+1}} - log\sum_{i=1}^{n}e^{g_i^t}$, $g(w^t) = [g_1, g_2, ..., g_n]$ denotes the logit output of the model, $\nabla f(w^t)$ is the gradient of the loss function $f(w^t)$, w^t denotes the parameters of model, h represents the learning rate and S is an irreducible constant. Please note that $max(g_i)$ here is not the common OOD score *MSP* (where $MSP = max(softmax(g_i))$), rather the maximum value of the output logits (without the softmax operation). The upper bound of the change of Energy score can effectively improve the separation between ID and OOD data.

For the detailed derivations of the upper bound, please refer to Sec. B of the supplementary material. We start by analyzing the *r.h.s* terms of Eq. (6), *i.e.,* the gradient norm and the

max logit. We can observe that gradient norm distributions are only mildly influenced, but the OOD data would decrease more in the max logit, making the distribution of the OOD scores decrease faster (within 50 training steps/mini-batches) and move more to the left. This would help the model to distinguish the ID and OOD samples. And the empirical observation meets our analysis of Prop. 1. This analysis clearly demonstrates why our proposed SFT could improve the OOD performance of the model. For the detailed explanation of the process and visualizations, please refer to Sec. A of the Supplementary Material.

B. A Machine Unlearning Perspective

Machine unlearning already has broad applications and extensive analytical discussions [29], [73]. We have found that the fine-tuning process of our SFT can be viewed as a special unlearning process. This insight motivates us to theoretically analyze our SFT from the perspective of machine unlearning.

In this part, we first introduce the fundamentals of robust machine unlearning. We then illustrate the similarity between our OOD fine-tuning and robust machine unlearning from the perspective of distribution shift and their training objectives, leading to similar challenges. Building on this, we establish a link between this dilemma and the optimization framework of SAM, which in turn allows us to explain the effectiveness of our SFT from the perspective of robust machine unlearning.

1) Preliminaries on robust machine unlearning.: The primary goal of standard machine unlearning is to efficiently remove the influence of particular data [32], [74]. The optimization objective is given by the following equation:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}_R} L_r(f(x),y) + \lambda \mathbb{E}_{(x',y')\sim\mathcal{D}_F} L_f(f(x'),y'). \quad (7)$$

where L_r and L_f are the retain and forget loss, $\mathcal{D}_R$ and $\mathcal{D}_F$ denote retain and forget training distribution, and $\lambda \geq 0$ is the hyperparameter that balances forgetting and retention.

Although variety of unlearning algorithms have been proposed by optimizing the standard objective function mentioned above [27], [28], [32], [75], robust machine unlearning still is difficult to achieve. [35]. Robust machine unlearning primarily addresses the threat of relearning attacks, which aim to recover forgotten knowledge by altering the weights of model [76]. The objective function for the relearning attack is given by:

$$\min_{\sigma} L_{relearn}(\theta + \sigma) \quad (8)$$

where $L_{relearn}$ is the relearning loss, and σ is the final optimized perturbation on original weight.

The goal of robust machine unlearning is to make the model robust against such relearning attacks. If the relearning objective $L_{relearn}$ is defined as $-L_f$, which mitigates the effects of the forgetting objective, we can derive that achieving robust machine unlearning requires the following min-max robust optimization by combining Eq. 7 and Eq. 8 [30]:

$$\min_{\theta} \max_{||\sigma||_p \leq \rho} \lambda L_f(\theta + \sigma | D_F) + L_r(\theta | D_R) \quad (9)$$

where $||.||_p$ is the l_p norm, and $L(\theta|D)$ denotes the loss under the distribution D with the weight θ. [30] demonstrates that utilizing the optimization objective in the form of Eq. 9 enhances the robustness of machine unlearning.

2) Bridging OOD Fine-tuning & Robust Machine Unlearning.: From the perspective of machine unlearning, we reconsider the OOD fine-tuning problem. We interpret OOD fine-tuning as a mechanism to induce forgetting of OOD-related knowledge in the model via fine-tuning, thus mitigating the overconfidence on OOD data. The OOD-related knowledge is comparable to harmful knowledge or privacy information in machine unlearning. The Fig. 5 depicts the similarity in objective of these two tasks from the standpoint of distribution shift. Both tasks aim to maximize the distance between the metric score distributions of two different kinds of data. Specifically, OOD fine-tuning seeks to making ID and OOD data more separable on the OOD score distribution. For machine unlearning, the objective is to make the retain and forget data more distinguishable on the distribution of MIA (Membership Inference Attacks) success rate [1].

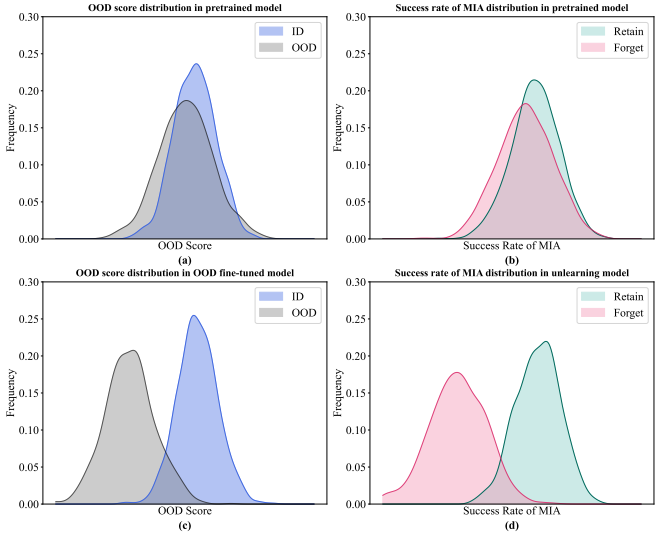


Fig. 5. Illustration of the similarity in objective of OOD Fine-tuning and Machine Unlearning. (a), (c) The OOD score distribution in the pretrained/fine-tuned model; (b), (d) The success rate of MIA distribution in the pretrained/unlearning model. From the perspective of distribution shift, these two tasks are remarkably similar.

Furthermore, the overall optimization objectives of these two tasks are extremely similar, presented in Eq. 5 and Eq. 7. Table I illustrates the correspondence between them:

1) From the standpoint of data sources, $\mathcal{D}_{\text{in}}$ and $\mathcal{D}_{\text{out}}$ in OOD fine-tuning are analogous to $\mathcal{D}_{\text{Retain}}$ and $\mathcal{D}_{\text{Forget}}$ in machine unlearning. $\mathcal{D}_{\text{out}}$ and $\mathcal{D}_{\text{Forget}}$ are the datasets designated for forgetting, while $\mathcal{D}_{\text{in}}$ and $\mathcal{D}_{\text{Retain}}$ are the datasets used to maintain performance.
2) With respect to the training loss functions, L_{CE} and L_{OOD} correspond to L_{retain} and L_{forget}. L_{OOD} and L_{forget} induce the model to forget specific knowledge, whereas L_{CE} and L_{retain} ensure the model preserves its original capabilities to avoid catastrophic forgetting.

Therefore, it is reasonable to consider OOD fine-tuning as a specific machine unlearning process.

[1]MIA success rate evaluates whether the model contains knowledge of the specific data. For retain data, the success rate should be high; for forget data, it should be approximately 50%, representing the random guess.

TABLE I
THE CORRESPONDENCE BETWEEN OOD FINE-TUNING AND MACHINE UNLEARNING.

Component	OOD fine-tuning	Machine Unlearning
Positive Samples	$\mathcal{D}_{\text{in}}$	$\mathcal{D}_{\text{Retain}}$
Negative Samples	$\mathcal{D}_{\text{out}}$	$\mathcal{D}_{\text{Forget}}$
Positive Objective	L_{CE}	L_{retain}
Negative Objective	L_{OOD}	L_{forget}

Similarly, achieving robust OOD detection is also a challenge for OOD fine-tuning. To illustrate, Granz et al. notes that perturbing model weights with random projections leads to similar scores for both ID and OOD data, making two kinds of data hard to separate [77]. As depicted in Fig. 5, the perturbation causes the OOD score distribution to shift from the pattern in Fig. 5 (c) back to the one in Fig. 5 (a). This phenomenon is very similar to the vulnerability of traditional unlearning methods in robust machine unlearning when faced with relearning attacks [30], [76]. As shown in the Fig. 5, the distribution of the MIA success rate reverts from the state in Fig. 5 (d) to that in Fig. 5 (b).

From the analysis above, it is evident that OOD fine-tuning and machine unlearning are highly similar in terms of both their fundamental and training objectives. Consequently, this allows us to naturally frame OOD detection as a specific instance of machine unlearning problem, thereby enabling an analysis from that perspective.

3) Our `SFT` leads to robust OOD detection.: As previously introduced, the optimization objective presented in Eq. 9 leads to more robust machine unlearning, and OOD fine-tuning can be viewed as a special case of machine unlearning process. Therefore, to achieve robust OOD detection, it is straightforward to transform the optimization objective of our `SFT` (Eq. 5) into the min-max robust optimization problem by following a procedure similar to the derivation of Eq. 9. The min-max robust optimization problem is shown in the following equation:

$$\min_{\theta} \underbrace{\max_{||\sigma||_p \leq \rho} \lambda L_{ood}(\theta+\sigma|D_{out})}_{l_{SAM}(\theta)} + L_{CE}(\theta|D_{in}) \tag{10}$$

where L_{ood} is the KL divergence between output and uniform distribution, and $l_{SAM}(\theta)$ denoted the SAM loss with $p = 2$ following the setting of Foret et al. [12].

The key insight from the above equation is that our `SFT` naturally requires the fine-tuning to achieve robust OOD detection by using SAM as the optimizer [2]. The experimental results in Table VII provides empirical support for our analysis. Additionally, since our analysis is not specific to any particular *post-hoc* OOD detection method, our `SFT` is theoretically compatible with multiple *post-hoc* methods. This is validated by the results in Fig. 6.

Through the analysis from these two perspectives, we theoretically demonstrate that our `SFT` not only enhances OOD detection performance, but also that SAM is particularly well-suited for OOD fine-tuning. Collectively, this establishes the solid theoretical foundation for our proposed `SFT`.

[2]Notably, we also employ SAM for the original ID training objective. This serves as a countermeasure to the ID accuracy drop often induced by OOD fine-tuning, given the proven benefits from SAM for ID performance [12], [18], [19]. As shown in Table XIII, our `SFT` successfully maintains or even enhances ID accuracy of the model.

V. EXPERIMENTS

A. Setup

Dataset. In line with Huang and Li [7], we use the large-scale ImageNet-1k benchmark [72] to evaluate our method, which is more challenging than the traditional benchmarks. We use four test sets as our OOD datasets, namely iNaturalist [83], SUN [84], Places [85], and Textures [86]. Besides the experiment on the large-scale benchmark, we also evaluate our method on CIFAR [87] benchmark where the OOD datasets consist of SVHN [88], Textures [86], LSUN-crop [89], LSUN-resize [89], iSUN [90] and Places365 [85].

Model. For the large-scale Imagenet-1k benchmark, we finish the main evaluation using Google BiT-S model [91] pretrained on ImageNet-1k with ResNetv2-101 [78]. Also, we evaluate our method with T2T-ViT-24 [92], which is a tokens-to-tokens vision transformer that exhibits competitive performance against CNNs when trained from scratch. Likewise, we verify the few-shot OOD detection performance with CLIP-B/16 [68] that achieves significant success in various downstream tasks by aligning images and raw texts using two separate encoders.

Evaluation Metrics. We mainly use FPR95 and AUROC as the evaluation metrics. FPR95 refers to the probability of misclassifying samples when the true positive rate (TPR) reaches 95%. Lower FPR95 implies less overlapping between ID and OOD scores. AUROC denotes the area under the receiver operating characteristic curve (ROC). Higher AUROC indicates better performance.

Implementation Details. Throughout all the experiments, we set the OOD loss strength λ of Eq. (5) as 1. For the pseudo-OOD data generated from *Jigsaw Puzzle Patch Shuffling*, we randomly select the patch size from $[2, 3, 5, 10, 15, 30, 60]$ at every iteration. Our fine-tuning lasts only for 50 steps with a learning rate of $1e{-}3$ and weight decay of $1e{-}5$. The perturbation radius ρ of SAM is set to 0.5. At the inference time, all images are resized to the resolution of $480{\times}480$.

B. Results

1) Main Results on ImageNet-1k.: Table II and Fig 6 compare the performance of some baseline methods and our fine-tuning method. We combine `SFT` with several *post hoc* methods and observe that our `SFT` can improve the performance of these *post hoc* methods by a large margin. Specifically, for `MSP` [13], `SFT` improves the baseline by 23% in the average FPR95 and 7.11% in the average AUROC. For `Energy` [6], `SFT` boosts the performance by 25.51% in the average FPR95 and 7.37% in the average AUROC. Equipped with our `SFT`, RankFeat [2] achieves the performance gain of 4.3% in the average FPR95 and 0.91% in the average AUROC. Powered by our `SFT`, `VRA` [15] attains performance improvement of 3.44% in the average FPR95 and 0.65% in the average AUROC. In particular, utilizing our `SFT`, the latest *state-of-the-art* `SCALE` [80] enhances the performance

TABLE II
MAIN EVALUATION RESULTS ON RESNETV2-101 [78]. THE BEST THREE RESULTS ARE HIGHLIGHTED WITH **RED**, **BLUE**, AND **CYAN**. ALL VALUES ARE REPORTED IN PERCENTAGES.

Methods	iNaturalist		SUN		Places		Textures		Average	
	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)
MSP [13]	63.69	87.59	79.89	78.34	81.44	76.76	82.73	74.45	76.96	79.29
ODIN [51]	62.69	89.36	71.67	83.92	76.27	80.67	81.31	76.30	72.99	82.56
Energy [6]	64.91	88.48	65.33	85.32	73.02	81.37	80.87	75.79	71.03	82.74
MaxLogit [79]	62.71	89.38	72.19	83.80	76.47	80.63	81.22	76.29	73.15	82.53
Mahalanobis [4]	96.34	46.33	88.43	65.20	89.75	64.46	52.23	72.10	81.69	62.02
GradNorm [5]	50.03	90.33	46.48	89.03	60.86	84.82	61.42	81.07	54.70	86.71
ReAct [56]	44.52	91.81	52.71	90.16	62.66	87.83	70.73	76.85	57.66	86.67
RankFeat [2]	41.31	91.91	**29.27**	**94.07**	**39.34**	**90.93**	37.29	91.70	36.80	92.15
ASH [14]	21.22	96.37	38.42	90.85	51.37	88.10	14.73	97.19	31.44	93.13
VRA [15]	20.81	97.70	32.89	92.68	45.83	90.01	23.88	95.43	30.85	93.71
SCALE [80]	17.53	96.98	35.12	91.90	47.29	89.44	15.78	96.91	28.93	93.81
ConjNorm [81]	50.54	89.84	63.39	81.36	74.62	76.84	29.33	93.91	54.47	85.49
KL Matching [8]	27.36	93.00	67.52	78.72	72.61	76.49	49.70	87.07	54.30	83.82
MOS [7]	**9.28**	**98.15**	40.63	92.01	49.54	89.06	60.43	81.23	39.97	90.11
ViM [82]	51.52	90.97	61.32	86.67	70.22	82.58	**4.90**	**98.92**	46.99	89.79
MSP+SFT	36.59	92.81	67.32	83.05	73.2	80.40	38.72	89.34	53.96	86.4
Energy+SFT	34.59	94.21	52.98	89.23	63.7	85.04	30.80	91.94	45.52	90.11
RankFeat+SFT	37.15	92.93	**27.66**	**94.35**	**37.88**	**91.14**	27.29	93.83	32.50	93.06
ASH+SFT	**13.78**	**97.49**	37.8	91.56	51.81	88.17	**8.53**	**98.33**	**27.98**	**93.89**
VRA+SFT	17.53	97.03	**33.19**	**93.13**	**47.12**	**89.67**	11.79	97.62	**27.41**	**94.36**
SCALE+SFT	**11.51**	**97.78**	34.68	92.45	48.33	89.33	**8.63**	**98.24**	**25.79**	**94.45**

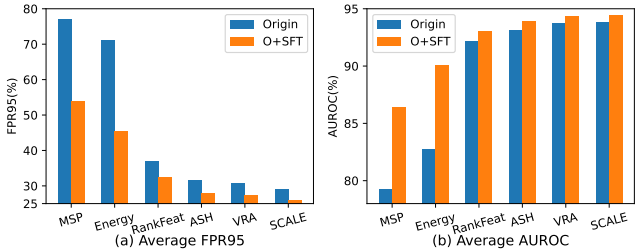


Fig. 6. Evaluation results on ResNetv2-101 [78] for various OOD detection methods. Powered by our SFT, all these methods show large improvements in both average FPR95 and AUROC, which demonstrates the robustness and versatility of our SFT.

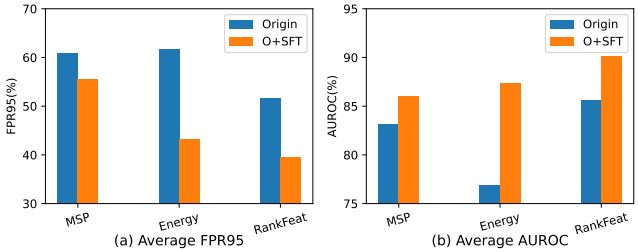


Fig. 7. Evaluation results on T2T-ViT-24 [92] for various OOD detection methods. By leveraging our SFT, all these methods achieve substantial improvements in both average FPR95 and AUROC, highlighting the robustness and adaptability of SFT.

by 3.14% in the average FPR95 and 0.64% in the average AUROC to achieve a new *state-of-the-art* performance. Compared to KL Matching [8], MOS [7] and ViM [82] that require a large amount of training time, our approach has better performance while taking much less training time. Moreover, the results from the Fig. 6 indicate that our SFT could boost the performance of all baseline methods. The consistent performance across various *post-hoc* methods demonstrates the robustness and versatility of our SFT.

2) Our SFT also suits architectures of different normalization layers.: Since our SAM-only-norm optimizer perturbs only the normalization layer, it is not clear whether the performance gain is consistent with other normalization layers. To investigate this question, we evaluate our method on T2T-ViT-24 [92] which uses LayerNorm throughout the network. Table III and Fig 7 compare the performance of some baseline methods as well as our SFT. For MSP, Energy, and RankFeat, our method brings the average performance gain of 9.10% in FPR95 and 5.95% in AUORC. The consistent improvement across methods demonstrates that our fine-tuning suits networks with different normalization layers. However, the ResNet50 with BatchNorm could not achieve the performance enhancement equipped with our SFT. We observe this intriguing behavior and conduct some experiments and analysis in Sec. VI.

3) Our SFT is also applicable to other benchmarks.: Now we turn to evaluating our approach on CIFAR-10 benchmark with VGG11 [93]. Table IV compares our method with some *post hoc* methods and training-needed methods. Similar to the ImageNet-1K benchmark, our SFT improves the Energy baseline by 19.59% in the average FPR95 and 3.53% in the average AUROC. Moreover, our SFT can boost the performance of the simple MSP to a very competitive level: 35.05% in the average FPR95 and 93.92% in the average AUROC.

4) Our SFT can be extended to CLIP in few-shot settings as well.: To demonstrate the validity of our SFT on CLIP, we evaluate our method on CLIP in a few-shot setting. Table V compares the performance of the CLIP-based OOD detection method, namely MCM [63], CoOp [69], LoCoOp [65], Dual-Adapter [71], NegPrompt [67], GalLoP [70], GOOD [95], and our SFT. Equipped with our SFT, MCM achieves a performance gain of 8.99% and 15.11% in the average FPR95; 1.96% and 2.46% in the average AUROC. Moreover, compared to CoOp [69], LoCoOp [65], Dual-Adapter [71], Neg-

TABLE III
EVALUATION RESULTS ON T2T-VIT-24 [92]. ALL VALUES ARE REPORTED IN PERCENTAGES.

Model	Methods	iNaturalist		SUN		Places		Textures		Average	
		FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)
T2T-ViT-24	**MSP** [13]	48.92	88.95	61.77	81.37	69.54	80.03	62.91	82.31	60.79	83.17
	ODIN [51]	44.07	88.17	63.83	78.46	68.19	75.33	54.27	83.63	57.59	81.40
	Mahalanobis [4]	90.50	58.13	91.71	50.52	93.32	49.60	80.67	64.06	89.05	55.58
	Energy [6]	52.95	82.93	68.55	73.06	74.24	68.17	51.05	83.25	61.70	76.85
	GradNorm [5]	99.30	25.86	98.37	28.06	99.01	25.71	92.68	38.80	97.34	29.61
	ReAct [56]	52.17	89.51	65.23	81.03	68.93	78.20	52.54	85.46	59.72	83.55
	RankFeat [2]	50.27	87.81	57.18	**84.33**	66.22	80.89	32.64	89.36	51.58	85.60
	VRA [15]	67.86	88.37	74.06	81.2	72.84	80.41	72.15	80.28	71.73	82.57
	MSP+SFT	41.98	90.97	65.41	82.88	68.03	**81.99**	47.02	88.12	55.61	85.99
	Energy+SFT	**30.90**	**92.13**	**56.02**	83.27	**63.29**	79.60	**22.27**	**94.56**	**43.12**	**87.39**
	RankFeat+SFT	**30.62**	**93.23**	**49.88**	**88.37**	**59.29**	**85.52**	**18.26**	**93.20**	**39.51**	**90.08**
	VRA+SFT	70.32	88.8	74.18	82.32	73.57	81.76	67.62	80.98	71.42	83.47

TABLE IV
EVALUATION RESULTS ON VGG11 [93]. ALL VALUES ARE REPORTED IN PERCENTAGES.

Model	Methods	SVHN		Textures		LSUN-crop		LSUN-resize		iSUN		Places365		Average	
		FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)
VGG11	**MSP** [13]	68.07	90.02	63.86	89.37	46.63	93.73	70.19	86.29	71.81	85.71	68.08	87.25	64.77	88.73
	ODIN [51]	53.84	92.23	48.09	91.94	19.95	97.01	54.29	89.47	56.61	88.87	52.34	89.86	47.52	91.56
	Energy [6]	53.13	92.26	47.04	92.08	18.51	97.20	53.02	89.58	55.39	88.97	51.67	89.95	46.46	91.67
	ReAct [56]	58.81	83.28	51.73	87.47	23.40	94.77	47.19	89.68	51.30	88.07	50.47	87.39	47.15	88.44
	VRA [15]	45.61	**93.58**	52.77	90.95	28.38	95.97	50.91	90.14	55.15	88.58	52.39	89.88	47.54	91.52
	DICE [94]	47.81	93.27	50.95	91.77	16.73	97.06	64.26	87.83	65.83	87.43	59.23	88.53	50.80	90.98
	FeatureNorm [10]	**8.84**	**98.24**	**24.62**	**95.11**	**3.38**	**99.36**	71.17	83.12	62.80	86.05	65.25	85.20	39.34	91.18
	MSP+SFT	57.15	91.10	54.68	91.03	9.00	98.44	15.29	97.13	18.27	96.65	55.41	89.16	35.05	93.92
	Energy+SFT	**44.23**	93.27	**43.39**	**92.90**	**3.25**	**99.16**	**10.65**	**97.86**	**13.51**	**97.36**	**46.18**	**90.64**	**26.87**	**95.20**
	VRA+SFT	48.86	92.58	45.69	92.44	5.2	98.94	**11.04**	**97.83**	**13.62**	**97.31**	**47.12**	**90.44**	**28.59**	**94.92**

Prompt [67], GalLoP [70], and GOOD [95] that require a large amount of training time, our approach has better performance while taking much less training time.

TABLE V
MAIN EVALUATION RESULTS REPORTED IN PERCENTAGES ON CLIP-B/16 [68].

Methods	iNaturalist		SUN		Places		Textures		Average	
	FPR-95(↓)	AUROC(↑)	FPR-95(↓)	AUROC(↑)	FPR-95(↓)	AUROC(↑)	FPR-95(↓)	AUROC(↑)	FPR-95(↓)	AUROC(↑)
Zero-shot										
MCM [13]	30.94	94.61	37.67	92.56	44.76	89.76	57.91	86.10	42.82	90.76
One-shot										
CoOp [69]	43.38	91.26	38.53	91.95	46.68	89.09	50.64	87.83	44.81	90.03
LoCoOp [65]	38.49	92.49	33.27	93.67	39.23	91.07	49.25	89.13	40.17	91.53
Dual-Adapter [71]	26.26	94.64	33.71	92.92	39.16	90.59	38.79	90.98	34.38	92.28
NegPrompt [67]	65.03	84.56	44.39	89.63	51.31	86.55	87.60	63.76	62.08	81.13
GalLoP [70]	31.63	93.61	36.05	92.59	43.60	90.03	53.71	86.83	41.25	90.77
GOOD [95]	28.60	94.10	33.44	93.23	41.57	90.26	49.77	88.84	38.34	91.60
SFT	33.35	93.78	34.65	92.88	44.24	89.55	23.09	94.67	**33.83**	**92.72**
16-shot										
CoOp [69]	28.00	94.43	36.95	92.29	43.03	89.74	39.33	91.24	36.83	91.93
LoCoOp [65]	23.06	95.45	32.70	93.35	39.92	90.64	40.23	91.32	33.98	92.69
Dual-Adapter [71]	23.33	95.22	25.89	94.44	34.08	91.55	38.72	91.45	30.50	93.62
NegPrompt [67]	37.79	90.49	32.11	92.25	35.52	91.16	43.93	88.38	37.34	90.57
GalLoP [70]	30.36	93.73	33.12	93.12	39.47	90.83	46.26	89.41	37.30	91.77
GOOD [95]	26.16	94.38	31.55	93.37	39.01	90.23	39.22	90.99	33.99	92.24
SFT	15.69	96.55	28.96	92.97	40.53	89.61	25.64	93.75	**27.71**	**93.22**

5) Our fine-tuning approach could enhance the robustness of OOD detectors: As [8], [96] point out, evaluating an OOD detector on a range of simple and synthetic OOD classes could reveal more about its OOD detection limitations. In order to further investigate the superiority of our method, we conduct the OOD unit-tests on the NINCO [96]. Table VI presents the OOD unit-test performance before and after our fine-tuning. After utilizing our approach, `Energy` could achieve the perfect performance in all the test units, which indicates that our method could boost the robustness of OOD detectors. For more details, please refer to the supplementary material.

TABLE VI
OOD UNIT-TEST RESULTS ON RESNETV2-101 [78].

Methods	Total test-units	Total failures(↓)	Pass rate(↑)
Energy	17	7	58.82%
Energy+SFT	17	**0**	**100%**

C. Ablation study

1) SAM is more effective than other optimizers.: To demonstrate the superiority of the SAM optimizer, we conduct an ablation study using alternative optimizers. Table VII compares the performance of SGD, Adam, and our SAM. Our SAM achieves better performance than the other baselines. Specifically, our SAM outperforms SGD by 4.72%, 4.76%, 4.99%, 4.75% and 5% in FPR95 with `MSP`, `MaxLogit`, `Energy`, `ODIN` and `GradNorm`, respectively. We attribute this improvement to the smooth loss landscape of ID samples optimized by SAM. And we conduct the same ablation study under the few-shot OOD detection setting. As can be seen from Table VIII, our SAM consistently achieves better performance compared to other optimizers, which further demonstrates the effectiveness of our SAM.

2) Comparison with OE [9]: To demonstrate the validity and efficiency of our `SFT`, we compare our method with

TABLE VII
EVALUATIONS RESULTS (FPR95↓) OF DIFFERENT OPTIMIZERS ON RESNETV2-101 [78]. ALL VALUES ARE REPORTED IN PERCENTAGES.

Methods	iNaturalist	SUN	Places	Textures	Average
MSP+FT(SGD)	44.18	73.54	76.63	40.35	58.68
MSP+FT(Adam)	65.65	82.97	81	55.85	71.37
MSP+SFT	**36.59**	**67.32**	**73.2**	**38.72**	**53.96**
MaxLogit+FT(SGD)	40.12	65.83	71.63	34.36	52.99
MaxLogit+FT(Adam)	69.56	77.65	76.21	51.03	68.61
MaxLogit+SFT	**33.33**	**59.4**	**67.35**	**32.85**	**48.23**
Energy+FT(SGD)	41.95	60.22	68.22	31.65	50.51
Energy+FT(Adam)	74.09	75.38	74.35	49.40	68.31
Energy+SFT	**34.59**	**52.98**	**63.70**	**30.80**	**45.52**
ODIN+FT(SGD)	40.12	65.82	71.63	34.36	52.98
ODIN+FT(Adam)	67.62	76.7	75.33	49.29	67.24
ODIN+SFT	**33.32**	**59.41**	**67.35**	**32.85**	**48.23**
GradNorm+FT(SGD)	27.24	49.3	62.13	**19.65**	39.58
GradNorm+FT(Adam)	56.19	69	71.7	33.87	57.69
GradNorm+SFT	**21.49**	**41.38**	**55.35**	20.09	**34.58**

TABLE VIII
EVALUATIONS RESULTS (FPR95↓) OF DIFFERENT OPTIMIZERS ON CLIP-B/16 [68] IN THE 16-SHOT SETTING. ALL VALUES ARE REPORTED IN PERCENTAGES.

Methods	iNaturalist	SUN	Places	Textures	Average
MCM+FT(SGD)	33.22	42.71	52.52	31.83	40.07
MCM+FT(Adam)	18.83	30.19	39.62	30.69	29.83
MCM+SFT	15.69	28.96	40.53	25.64	**27.71**

the strong baseline OE. Table IX compares the performance of our SFT and OE on VGG11 [93]. As can be seen, our SFT achieves the performance on par with OE, which indicates the efficacy of SFT. Besides, Table X shows the needed fine-tuning epochs/steps and time cost of SFT and OE. Compared with OE, our SFT demonstrates notably reduced time consumption. Moreover, SFT not requiring additional real OOD data enhances its potential for practical deployment.

3) Comparison with the OE-based (C+1) classifier.: We trained an OE-based (C+1) classifier and present the results in Table XI. Our SFT outperforms the OE-based (C+1) classifier by 10.77% in the average of FPR95 and 1.58% in the average of AUROC within much fewer training steps.

4) Real OOD data and methods.: Table XII presents the evaluation results of our SFT with real OOD data and OE on the CIFAR benchmark. Compared with our original SFT, using real OOD data improves the performance by 5.99% in FPR95 and 0.82% in AUROC. Our method also outperforms OE which uses real OOD data. The result confirms the effectiveness of our SFT when the real OOD data is accessible. However, real OOD data are expensive to annotate and access in real-world applications. In contrast, our SFT is OOD data-free and thus more practical.

5) SAM-only-norm instead of SAM.: We initially used the original SAM as the optimizer but we observed that the original SAM would severely decrease the standard classification accuracy of the pre-trained model. We expect this might be related to the strong perturbation of SAM applied at every layer. Table XIII compares the performance of SAM and SAM-only-norm on different architectures. As can be seen, SAM-only-norm can simultaneously improve the OOD detection performance while maintaining the standard classification accuracy. We thus switch to SAM-only-norm in the experiments. We stress that this performance, achieved by our SFT, is crucial for OOD detection, as any effective OOD detector must not degrade ID accuracy. Moreover, it would be easy to conduct theoretical analysis for SAM-only-norm as only the normalization layers are involved.

6) Impact of perturbation radius ρ.: One important hyperparameter of SAM is the perturbation radius ρ. A large ρ might result in a flat loss landscape for both ID and OOD data, while a small ρ might have no impact of distinguishing OOD samples. To choose an appropriate ρ, we conduct an ablation and present the results in Table XIV. As can be seen, $\rho = 0.5$ achieves the best performance among different values. We note that though our SFT needs to search for an appropriate ρ, it is still very efficient as the fine-tuning only takes 50 steps.

7) Impact of pseudo-OOD data.: Our SFT mainly consists of pseudo-data generation and SAM fine-tuning. To validate the effectiveness of these main components, we conducted the ablation studies in these two components. Table XV compares the performance of different components. As can be seen, our SFT, which combines pseudo-data generation and SAM fine-tuning, achieves the best OOD detection performance. This validates the effectiveness of the two core components in our method.

8) Ablation on different pseudo-OOD data: The pseudo-OOD data plays an important role in our SFT. Table XVI compares the performance of different generation methods. The generation method combining *Jigsaw Puzzle Patch Shuffling* and *RGB Channel Shuffling* obtains the best performance, which demonstrates the efficacy of our generation of the pseudo-OOD data.

9) Domain Generalization : To further demonstrate the generality of our SFT, we conducted the domain generalization evaluations on ImageNet-V2 [97] and ImageNet-R [98]. As Fig. 8 indicates, our SFT enhanced the domain generalization ability of the pretrained model across various test datasets. These results demonstrate that our SFT not only enhances OOD detection capability but also improves domain adaption performance, further confirming the effectiveness and robustness of our approach.

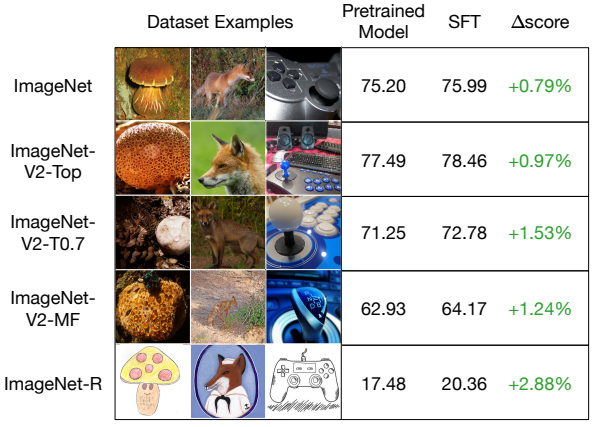


Fig. 8. Visualizing distribution shift of different domain test datasets. The domain generalization performance of pretrained model and our SFT. All values are reported in percentages.

TABLE IX
EVALUATION RESULTS OF OUR SFT AND OE [9] ON VGG11 [93]. ALL VALUES ARE REPORTED IN PERCENTAGES.

Methods	SVHN		Textures		LSUN-crop		LSUN-resize		iSUN		Places365		Average	
	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)
Energy+OE [9]	13.60	97.32	18.39	96.43	12.19	97.51	29.6	94.62	24.62	95.39	36.02	92.80	22.4	95.68
Energy+SFT	44.23	93.27	43.39	92.90	3.25	99.16	10.65	97.86	13.51	97.36	46.18	90.64	26.87	95.20

TABLE X
COMPARISON OF COMPUTATIONAL EFFICIENCY ON VGG11 [93].

Methods	fine-tuning epochs or steps	time cost(s)
OE	100 epochs	847.65
SFT	300 steps	19.40

TABLE XI
EVALUATION RESULTS ON VGG11 AVERAGED ON SVHN, TEXTURES, LSUN-CROP, LSUN-RESIZE, ISUN, AND PLACES365. OE(C+1) DENOTES THAT WE TRAIN AN OE-BASED (C+1) CLASSIFIER FOR OOD DETECTION. ALL VALUES ARE REPORTED IN PERCENTAGES.

Methods	Fine-tuning Steps	FPR95 (↓)	AUROC (↑)
Energy	0	46.46	91.67
SFT	**300 steps**	**26.87**	**95.20**
OE(C+1)	50 epochs	37.64	93.62

VI. LIMITATION AND DISCUSSION

A. More diverse and realistic pseudo-OOD data are needed.

As Fig. 3 and Table II indicate, our SFT has very large improvement on Textures and iNaturalist. This observation indicates that our pseudo-OOD data might precisely represent the image patterns of Textures and iNaturalist. However, the performance improvement is relatively limited on Places and SUN, which indicates the pseudo-OOD are not able to imitate the images of these two datasets. Moreover, in the ablation study, we show that fine-tuning for long would harm the performance. This phenomenon might be also related to the diversity of the generated pseudo-OOD data. Seeking a more powerful way to generate diverse and realistic pseudo-OOD data is an important future research direction. Also, we additionally employ Mosaic and Speckle to generate pseudo-OOD data. As can be seen from Table XVIII, Mosaic and Speckle fail to improve the performance further. This indicates that it is challenging to seek a more effective and diverse way to generate pseudo-OOD data.

TABLE XII
EVALUATION RESULTS ON VGG11 AVERAGED ON SVHN, TEXTURES, LSUN-CROP, LSUN-RESIZE, ISUN, AND PLACES365. ALL VALUES ARE REPORTED IN PERCENTAGES.

Methods	real OOD free?	FPR95 (↓)	AUROC (↑)
Energy+SFT	✓	26.87	95.20
Energy+OE	✗	22.4	95.68
Energy+SFT (real OOD data)	✗	**20.88**	**96.02**

TABLE XIII
ABLATION STUDY OF SAM AND SAM-ONLY-NORM FINE-TUNING ON RESNETV2-101 [78] AND T2T-VIT-24 [92]. HERE ↓ / ↑ MEANS THE MAGNITUDE OF THE ACCURACY WOULD DECREASE / INCREASE COMPARED TO THE PRE-TRAINED MODEL. ALL VALUES ARE REPORTED IN PERCENTAGES.

Model	Optimizer	FPR95(↓)	AUROC(↑)	ACC (%)
ResNetv2-101	SAM	28.57	93.35	70.01 (5.19 ↓)
	SAM-only-norm	45.52	90.11	75.99 (0.79 ↑)
T2T-ViT-24	SAM	93.73	48.38	0.83 (80.69 ↓)
	SAM-only-norm	43.12	87.39	80.74 (0.78 ↓)

TABLE XIV
SENSITIVITY ANALYSIS OF THE HYPERPARAMETER ρ ON RESNETV2-101 [78]. ALL VALUES ARE REPORTED IN PERCENTAGES.

ρ	0.05	0.1	0.5	1.0	2.0
FPR95 (↓)	46.62	45.81	**45.52**	46.02	47.92
AUROC (↑)	89.74	89.93	**90.11**	89.93	89.31

TABLE XV
ABLATION STUDY ON THE IMPACT OF PSEUDO-OOD DATA DURING THE FINE-TUNING ON RESNETV2-101 [78] AND T2T-VIT-24 [92]. ALL VALUES ARE REPORTED IN PERCENTAGES. SFT-ID DENOTES SFT ONLY UTILIZING ID DATA. FT(SGD)-ID DENOTES FINE-TUNING THE MODEL WITH SGD ONLY UTILIZING ID DATA, AND FT(SGD) REPRESENTS FINE-TUNING THE MODEL WITH SGD.

Model	Methods	Average		Methods	Average	
		FPR95 (↓)	AUROC (↑)		FPR95 (↓)	AUROC (↑)
ResNetv2-101	Energy	71.03	82.74	Energy	71.03	82.74
	SFT-ID	71.42	83.23	FT(SGD)-ID	72.23	82.8
	SFT	**45.52**	**90.11**	FT(SGD)	**50.51**	**88.84**
T2T-ViT-24	Energy	61.70	76.85	Energy	61.70	76.85
	SFT-ID	61.45	78.04	FT(SGD)-ID	61.33	77.54
	SFT	**43.12**	**87.39**	FT(SGD)	**48.29**	**84.8**

B. BatchNorm does not suit our SFT.

As mentioned in Sec. V, we observe that the architectures with BatchNorm usually could not achieve competitive performance compared with other normalization layers, such as GroupNorm and LayerNorm. To further delve into the distinction, we conduct the experiment of ResNet50 with

TABLE XVI
ABLATION STUDY ON THE GENERATION OF PSEUDO-OOD DATA DURING THE FINE-TUING ON RESNETV2-101 [78] AND T2T-VIT-24 [92]. ALL VALUES ARE REPORTED IN PERCENTAGES. *PS* DENOTES *Patch Shuffling*, & *CS* DENOTES *RGB Channel Shuffling*.

Model	Methods	FPR95 (↓)	AUROC (↑)
ResNetv2-101	SFT(only *PS*)	49.56	88.97
	SFT(only *CS*)	73.27	82.92
	SFT	**45.52**	**90.11**
T2T-ViT-24	SFT(only *PS*)	47.39	86.32
	SFT(only *CS*)	71.74	76.30
	SFT	**43.12**	**87.39**

TABLE XVII
MAIN EVALUATION RESULTS ON RESNET50 [99] WITH BATCHNORM AND GROUPNORM. ALL VALUES ARE REPORTED IN PERCENTAGES, HERE ↓ / ↑ MEANS THE MAGNITUDE OF THE OOD DETECTION WOULD DECREASE / INCREASE COMPARED TO THE PRETRAINED MODEL.

Model	Methods	iNaturalist		SUN		Places		Textures		Average	
		FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)	FPR95 (↓)	AUROC (↑)
ResNet50-BN(75.03%)	Energy	51.98	90.42	58.64	86.61	64.83	84.03	55.60	85.72	57.76	86.7
	VRA	13.01	97.47	23.47	94.86	33.99	92.00	31.88	93.44	25.59	94.44
	Energy+SFT	76.65	76.89	79.93	68.98	80.47	68.54	77.75	65.91	78.7 (20.94 ↓)	70.08 (16.62 ↓)
	VRA+SFT	80.32	80.44	83.63	70.44	83.19	70.19	63.34	79.33	77.62 (52.03 ↓)	75.1 (19.34 ↓)
ResNet50-GN(74.17%)	Energy	54.16	90.27	56.58	88.03	60.17	85.89	54.29	87.63	56.3	87.96
	VRA	42.78	91.72	56.38	86.95	59.97	84.9	65.02	79.68	56.04	85.81
	Energy+SFT	32.16	93.68	51.65	88.97	59.61	85.73	29.40	93.19	43.21 (13.09 ↑)	90.39 (2.43 ↑)
	VRA+SFT	24.43	94.83	46.12	89.58	54.22	86.40	26.72	93.35	37.87 (18.17 ↑)	91.04 (5.23 ↑)

TABLE XVIII
EVALUATION RESULTS ON RESNETV2-101.

Methods	**SFT**	**SFT+** Mosaic	**SFT+** Speckle	**SFT+** Mosaic &Speckle
FPR95 (↓)	**45.52**	46.28	47.36	58.68
AUROC (↑)	**90.11**	90.07	89.82	85.16

BatchNorm and GroupNorm. Firstly, we train the ResNet50-BN and ResNet50-GN from scratch with consistent training parameters. Table XVII presents the OOD detection performance of BatchNorm and GroupNorm. The consistent classification accuracy indicates that it is fair to analyze the contrasts between normalization layers using these two pretrained models [1]. Our SFT hurts the OOD detection performance of ResNet50 with BatchNorm, but improves the performance of ResNet50 with GroupNorm. This result provides evidence supporting our earlier observations. Implementing specialized adaptations to BatchNorm to enhance its compatibility with OOD fine-tuning is worth exploring in the future work.

VII. CONCLUSION

In this paper, we propose SFT, an OOD-data-free and time-efficient OOD detection method by fine-tuning the pre-trained model with SAM. Our SFT helps existing OOD detection methods to have large performance gains across various benchmarks. Extensive ablation studies are performed to illustrate different hyper-parameters, and theoretical analyses are presented to shed light on the working mechanism. We hope our method can bring inspiration and novel understandings to the community of OOD detection from alternative perspectives.

REFERENCES

[1] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Open-set recognition: A good closed-set classifier is all you need?" *arXiv preprint arXiv:2110.06207*, 2021.

[2] Y. Song, N. Sebe, and W. Wang, "Rankfeat: Rank-1 feature removal for out-of-distribution detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 885–17 898, 2022.

[3] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *CVPR*, 2015, pp. 427–436.

[4] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," *NeurIPS*, vol. 31, 2018.

[5] R. Huang, A. Geng, and Y. Li, "On the importance of gradients for detecting distributional shifts in the wild," *NeurIPS*, vol. 34, 2021.

[6] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," *NeurIPS*, vol. 33, pp. 21 464–21 475, 2020.

[7] R. Huang and Y. Li, "Mos: Towards scaling out-of-distribution detection for large semantic space," in *CVPR*, 2021, pp. 8710–8719.

[8] D. Hendrycks, S. Basart, M. Mazeika, M. Mostajabi, J. Steinhardt, and D. Song, "Scaling out-of-distribution detection for real-world settings," *ICML*, 2022.

[9] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," *arXiv preprint arXiv:1812.04606*, 2018.

[10] Y. Yu, S. Shin, S. Lee, C. Jun, and K. Lee, "Block selection method for using feature norm in out-of-distribution detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 701–15 711.

[11] X. Du, Y. Sun, J. Zhu, and Y. Li, "Dream the impossible: Outlier imagination with diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[12] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," in *International Conference on Learning Representations*, 2020.

[13] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *ICLR*, 2017.

[14] A. Djurisic, N. Bozanic, A. Ashok, and R. Liu, "Extremely simple activation shaping for out-of-distribution detection," in *The Eleventh International Conference on Learning Representations*, 2023.

[15] M. Xu, Z. Lian, B. Liu, and J. Tao, "Vra: Variational rectifed activation for out-of-distribution detection," *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[16] D. Bahri, H. Mobahi, and Y. Tay, "Sharpness-aware minimization improves language model generalization," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 7360–7371.

[17] X. Chen, C.-J. Hsieh, and B. Gong, "When vision transformers outperform resnets without pre-training or strong data augmentations," in *International Conference on Learning Representations*, 2021.

[18] J. Kwon, J. Kim, H. Park, and I. K. Choi, "Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5905–5914.

[19] J. Zhuang, B. Gong, L. Yuan, Y. Cui, H. Adam, N. C. Dvornek, J. s Duncan, T. Liu *et al.*, "Surrogate gap minimization improves sharpness-aware training," in *International Conference on Learning Representations*, 2021.

[20] M. Mueller, T. Vlaar, D. Rolnick, and M. Hein, "Normalization layers are all that sharpness-aware minimization needs," *arXiv preprint arXiv:2306.04226*, 2023.

[21] P. Mi, L. Shen, T. Ren, Y. Zhou, X. Sun, R. Ji, and D. Tao, "Make sharpness-aware minimization stronger: A sparsified perturbation approach," *Advances in Neural Information Processing Systems*, vol. 35, pp. 30 950–30 962, 2022.

[22] Y. Liu, S. Mai, X. Chen, C.-J. Hsieh, and Y. You, "Towards efficient and scalable sharpness-aware minimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 360–12 370.

[23] J. Du, H. Yan, J. Feng, J. T. Zhou, L. Zhen, R. S. M. Goh, and V. Tan, "Efficient sharpness-aware minimization for improved training of neural networks," in *International Conference on Learning Representations*, 2021.

[24] M. Andriushchenko, D. Bahri, H. Mobahi, and N. Flammarion, "Sharpness-aware minimization leads to low-rank features," *arXiv preprint arXiv:2305.16292*, 2023.

[25] M. Andriushchenko and N. Flammarion, "Towards understanding sharpness-aware minimization," in *International Conference on Machine Learning*. PMLR, 2022, pp. 639–668.

[26] A. Ginart, M. Guan, G. Valiant, and J. Y. Zou, "Making ai forget you: Data deletion in machine learning," *Advances in neural information processing systems*, vol. 32, 2019.

[27] M. Chen, W. Gao, G. Liu, K. Peng, and C. Wang, "Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7766–7775.

[28] C. Fan, J. Liu, Y. Zhang, E. Wong, D. Wei, and S. Liu, "Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation," *arXiv preprint arXiv:2310.12508*, 2023.

[29] J. Jang, D. Yoon, S. Yang, S. Cha, M. Lee, L. Logeswaran, and M. Seo, "Knowledge unlearning for mitigating privacy risks in language models," in *61st Annual Meeting of the Association for Computational Linguistics, ACL 2023*. Association for Computational Linguistics (ACL), 2023, pp. 14 389–14 408.

[30] C. Fan, J. Jia, Y. Zhang, A. Ramakrishna, M. Hong, and S. Liu, "Towards llm unlearning resilient to relearning attacks: A sharpness-aware minimization perspective and beyond," in *Forty-second International Conference on Machine Learning*, 2025.

[31] R. Zhang, L. Lin, Y. Bai, and S. Mei, "Negative preference optimization: From catastrophic collapse to effective unlearning," *arXiv preprint arXiv:2404.05868*, 2024.

[32] S. Liu, Y. Yao, J. Jia, S. Casper, N. Baracaldo, P. Hase, Y. Yao, C. Y. Liu, X. Xu, H. Li *et al.*, "Rethinking machine unlearning for large language models," *Nature Machine Intelligence*, pp. 1–14, 2025.

[33] C. Liu, Y. Wang, J. Flanigan, and Y. Liu, "Large language model unlearning via embedding-corrupted prompts," *Advances in Neural Information Processing Systems*, vol. 37, pp. 118 198–118 266, 2024.

[34] P. Thaker, Y. Maurya, S. Hu, Z. S. Wu, and V. Smith, "Guardrail baselines for unlearning in llms," *arXiv preprint arXiv:2403.03329*, 2024.

[35] F. Barez, T. Fu, A. Prabhu, S. Casper, A. Sanyal, A. Bibi, A. O'Gara, R. Kirk, B. Bucknall, T. Fist *et al.*, "Open problems in machine unlearning for ai safety," *arXiv preprint arXiv:2501.04952*, 2025.

[36] Z. Zhang, F. Wang, X. Li, Z. Wu, X. Tang, H. Liu, Q. He, W. Yin, and S. Wang, "Does your llm truly unlearn? an embarrassingly simple approach to recover unlearned knowledge," *arXiv e-prints*, pp. arXiv–2410, 2024.

[37] A. Dayal, S. Shrusti, L. R. Cenkeramaddi, C. K. Mohan, and A. Kumar, "Leveraging mixture alignment for multi-source domain adaptation," *IEEE Transactions on Image Processing*, 2025.

[38] M. Qi, P. Zhu, X. Li, X. Bi, L. Qi, H. Ma, and M.-H. Yang, "Dc-sam: In-context segment anything in images and videos via dual consistency," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[39] M. Lan, M. Meng, J. Yu, and J. Wu, "Learning to discover knowledge: a weakly-supervised partial domain adaptation approach," *IEEE Transactions on Image Processing*, vol. 33, pp. 4090–4103, 2024.

[40] M. Qi, H. Ye, J. Peng, and H. Ma, "Action quality assessment via hierarchical pose-guided multi-stage contrastive regression," *IEEE Transactions on Image Processing*, 2025.

[41] Z. Deng, K. Zhou, D. Li, J. He, Y.-Z. Song, and T. Xiang, "Dynamic instance domain adaptation," *IEEE Transactions on Image Processing*, vol. 31, pp. 4585–4597, 2022.

[42] B. Xu, Z. Zeng, C. Lian, and Z. Ding, "Few-shot domain adaptation via mixup optimal transport," *IEEE Transactions on Image Processing*, vol. 31, pp. 2518–2528, 2022.

[43] M. Qi, C. Lv, and H. Ma, "Robust disentangled counterfactual learning for physical audiovisual commonsense reasoning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[44] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *ICLR*, 2019.

[45] X. Liu, Y. Huang, H. Wang, Z. Xiao, and S. Zhang, "Universal and scalable weakly-supervised domain adaptation," *IEEE Transactions on Image Processing*, vol. 33, pp. 1313–1325, 2024.

[46] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," *NeurIPS*, vol. 32, 2019.

[47] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," in *ICLR*, 2019.

[48] Y.-C. Hsu, Y. Shen, H. Jin, and Z. Kira, "Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data," in *CVPR*, 2020, pp. 10 951–10 960.

[49] J. Serrà, D. Álvarez, V. Gómez, O. Slizovskaia, J. F. Núñez, and J. Luque, "Input complexity and out-of-distribution detection with likelihood-based generative models," in *ICLR*, 2019.

[50] P. Chong, N.-M. Cheung, Y. Elovici, and A. Binder, "Toward scalable and unified example-based explanation and outlier detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 525–540, 2021.

[51] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," *ICLR*, 2018.

[52] H. Zhao, H. Liu, Z. Ding, and Y. Fu, "Consensus regularized multi-view outlier detection," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 236–248, 2017.

[53] Y. Sun, Y. Ming, X. Zhu, and Y. Li, "Out-of-distribution detection with deep nearest neighbors," in *International Conference on Machine Learning*. PMLR, 2022, pp. 20 827–20 840.

[54] J. Ren, J. Luo, Y. Zhao, K. Krishna, M. Saleh, B. Lakshminarayanan, and P. J. Liu, "Out-of-distribution detection and selective generation for conditional language models," in *The Eleventh International Conference on Learning Representations*, 2022.

[55] M. S. Graham, W. H. Pinaya, P.-D. Tudosiu, P. Nachev, S. Ourselin, and J. Cardoso, "Denoising diffusion models for out-of-distribution detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2947–2956.

[56] Y. Sun, C. Guo, and Y. Li, "React: Out-of-distribution detection with rectified activations," *NeurIPS*, vol. 34, 2021.

[57] Q. Wu, Y. Chen, C. Yang, and J. Yan, "Energy-based out-of-distribution detection for graph neural networks," in *The Eleventh International Conference on Learning Representations*, 2022.

[58] Y. Song, W. Wang, and N. Sebe, "Rankfeat&rankweight: Rank-1 feature/weight removal for out-of-distribution detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[59] Y. Wen, D. Tran, and J. Ba, "Batchensemble: an alternative approach to efficient ensemble and lifelong learning," in *International Conference on Learning Representations*, 2019.

[60] Y. Ming, Y. Fan, and Y. Li, "Poem: Out-of-distribution detection with posterior sampling," in *International Conference on Machine Learning*. PMLR, 2022, pp. 15 650–15 665.

[61] J. Zhang, L. Gao, B. Hao, H. Huang, J. Song, and H. Shen, "From global to local: Multi-scale out-of-distribution detection," *IEEE Transactions on Image Processing*, 2023.

[62] A. Wu, C. Deng, and W. Liu, "Unsupervised out-of-distribution object detection via pca-driven dynamic prototype enhancement," *IEEE Transactions on Image Processing*, 2024.

[63] Y. Ming, Z. Cai, J. Gu, Y. Sun, W. Li, and Y. Li, "Delving into out-of-distribution detection with vision-language representations," *Advances in Neural Information Processing Systems*, vol. 35, pp. 35 087–35 102, 2022.

[64] S. Esmaeilpour, B. Liu, E. Robertson, and L. Shu, "Zero-shot out-of-distribution detection based on the pre-trained model clip," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 6, 2022, pp. 6568–6576.

[65] A. Miyai, Q. Yu, G. Irie, and K. Aizawa, "Locoop: Few-shot out-of-distribution detection via prompt learning," in *Thirty-Seventh Conference on Neural Information Processing Systems*, 2023.

[66] J. Dong, Y. Yao, W. Jin, H. Zhou, Y. Gao, and Z. Fang, "Enhancing few-shot out-of-distribution detection with pre-trained model features," *IEEE Transactions on Image Processing*, 2024.

[67] T. Li, G. Pang, X. Bai, W. Miao, and J. Zheng, "Learning transferable negative prompts for out-of-distribution detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 584–17 594.

[68] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[69] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[70] M. Lafon, E. Ramzi, C. Rambour, N. Audebert, and N. Thome, "Gallop: Learning global and local prompts for vision-language models," in *European Conference on Computer Vision*. Springer, 2024, pp. 264–282.

[71] X. Chen, Y. Li, and H. Chen, "Dual-adapter: Training-free dual adaptation for few-shot out-of-distribution detection," *arXiv preprint arXiv:2405.16146*, 2024.

[72] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. Ieee, 2009, pp. 248–255.

[73] E. Ullah, T. Mai, A. Rao, R. A. Rossi, and R. Arora, "Machine unlearning via algorithmic stability," in *Conference on Learning Theory*. PMLR, 2021, pp. 4126–4142.

[74] Y. Cao and J. Yang, "Towards making systems forget with machine unlearning," in *2015 IEEE symposium on security and privacy*. IEEE, 2015, pp. 463–480.

[75] M. Kurmanji, P. Triantafillou, J. Hayes, and E. Triantafillou, "Towards unbounded machine unlearning," *Advances in neural information processing systems*, vol. 36, pp. 1957–1987, 2023.

[76] S. Hu, Y. Fu, S. Wu, and V. Smith, "Jogging the memory of unlearned models through targeted relearning attacks," in *ICML 2024 Workshop on Foundation Models in the Wild*, 2024.

[77] M. Granz, M. Heurich, and T. Landgraf, "Weiper: Ood detection using weight perturbations of class projections," *Advances in Neural Information Processing Systems*, vol. 37, pp. 35 879–35 908, 2024.

[78] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *ECCV*. Springer, 2016, pp. 630–645.

[79] D. Hendrycks, S. Basart, M. Mazeika, A. Zou, J. Kwon, M. Mostajabi, and J. Steinhardt, "Improving and assessing anomaly detectors for large-scale settings," 2022. [Online]. Available: https://openreview.net/forum?id=vruwp11pWnO

[80] K. Xu, R. Chen, G. Franchi, and A. Yao, "Scaling for training time and post-hoc out-of-distribution detection enhancement," in *The Twelfth International Conference on Learning Representations*, 2024.

[81] B. Peng, Y. Luo, Y. Zhang, Y. Li, and Z. Fang, "Conjnorm: Tractable density estimation for out-of-distribution detection," in *The Twelfth International Conference on Learning Representations*, 2024.

[82] H. Wang, Z. Li, L. Feng, and W. Zhang, "Vim: Out-of-distribution with virtual-logit matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[83] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *CVPR*, 2018.

[84] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *CVPR*, 2010.

[85] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE TPAMI*, 2017.

[86] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *CVPR*, 2014.

[87] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[88] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," 2011.

[89] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.

[90] P. Xu, K. A. Ehinger, Y. Zhang, A. Finkelstein, S. R. Kulkarni, and J. Xiao, "Turkergaze: Crowdsourcing saliency with webcam based eye tracking," *arXiv preprint arXiv:1504.06755*, 2015.

[91] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (bit): General visual representation learning," in *ECCV*, 2020.

[92] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *ICCV*, 2021.

[93] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, 2015.

[94] Y. Sun and Y. Li, "Dice: Leveraging sparsification for out-of-distribution detection," in *European Conference on Computer Vision*. Springer, 2022, pp. 691–708.

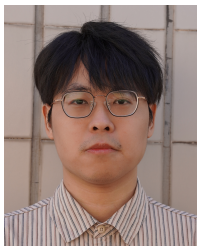
[95] P. Li, B. Cao, C. Zhang, and Q. Hu, "Generalized few-shot out-of-distribution detection," *arXiv preprint arXiv:2508.05732*, 2025.

[96] J. Bitterwolf, M. Müller, and M. Hein, "In or out? fixing imagenet out-of-distribution detection evaluation," in *International Conference on Machine Learning*. PMLR, 2023, pp. 2471–2506.

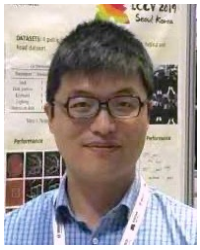
[97] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *International conference on machine learning*. PMLR, 2019, pp. 5389–5400.

[98] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo *et al.*, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 8340–8349.

[99] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.



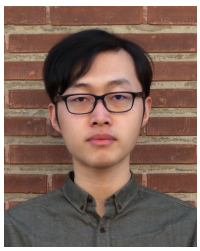
Chi Zhang received the B.S. degree and Master Degree from the China University of Mining Technology - Beijing. He is currently a Ph.D student at Beijing Jiaotong University. His research interests include computer vision, numerical optimization and deep learning.



Wei Wang (Member, IEEE) received the PhD degree from the University of Trento, in 2018. He is a full Professor at the Institute of Information Science at Beijing Jiaotong University since 2022. Before that, he was an Assistant Professor at the University of Trento and Postdoc at EPFL, Switzerland. His research interests include machine learning and its application to computer vision and multimedia analysis.

Yao Zhao (Fellow, IEEE) received the B.S. degree from the Radio Engineering Department, Fuzhou University, Fuzhou, China, in 1989, the M.E. degree from the Radio Engineering Department, Southeast University, Nanjing, China, in 1992, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University (BJTU), Beijing, China, in 1996. He is currently the Director of the Institute of Information Science, Beijing Jiaotong University. His current research interests include image/video coding and video analysis and understanding.

Nicu Sebe (Senior Member, IEEE) is Professor with the University of Trento, Italy, leading the Multimedia and Human Understanding Group. He was the General Co-Chair of ACM Multimedia 2013 and 2022, and the Program Chair of ACM Multimedia 2007 and 2011, ICCV 2017, ECCV 2016 and ICPR 2020. He is a fellow of the International Association for Pattern Recognition.



Yue Song received the B.Sc. *cum laude* from KU Leuven, Belgium and the joint M.Sc. *summa cum laude* from the University of Trento, Italy and KTH Royal Institute of Technology, Sweden, and the Ph.D. *summa cum laude* from the Multimedia and Human Understanding Group (MHUG) at the University of Trento, Italy. He was a post-doctoral research associate at Caltech, and is a assitant professor at the College of AI in Tsinghua University. His research interests are structured representation learning at the science – AI interface.