Unleashing Mask: Explore the Intrinsic Out-of-Distribution Detection Capability

Jianing Zhu [1] **Hengzhuang Li** [1] **Jiangchao Yao** [2 3] **Tongliang Liu** [4 5] **Jianliang Xu** [1] **Bo Han** [1]

Abstract

Out-of-distribution (OOD) detection is an indispensable aspect of secure AI when deploying machine learning models in real-world applications. Previous paradigms either explore better scoring functions or utilize the knowledge of outliers to equip the models with the ability of OOD detection. However, few of them pay attention to the intrinsic OOD detection capability of the given model. In this work, we generally discover the existence of an intermediate stage of a model trained on in-distribution (ID) data having higher OOD detection performance than that of its final stage across different settings, and further identify one critical data-level attribution to be learning with the *atypical samples*. Based on such insights, we propose a novel method, *Unleashing Mask*, which aims to restore the OOD discriminative capabilities of the well-trained model with ID data. Our method utilizes a mask to figure out the memorized *atypical samples*, and then finetune the model or prune it with the introduced mask to forget them. Extensive experiments and analysis demonstrate the effectiveness of our method. The code is available at: `https://github.com/tmlr-group/Unleashing-Mask`.

1. Introduction

Out-of-distribution (OOD) detection has drawn increasing attention when deploying machine learning models into the open-world scenarios (Nguyen et al., 2015; Lee et al., 2018a; Yang et al., 2021). Since the test samples can naturally arise from a label-different distribution, identifying OOD inputs from in-distribution (ID) data is important, especially for those safety-critical applications like autonomous driving and medical intelligence. Previous studies focus on designing a series of scoring functions (Hendrycks & Gimpel, 2017; Liang et al., 2018; Liu et al., 2020; Sun et al., 2022) for OOD uncertainty estimation or fine-tuning with auxiliary outlier data to better distinguish the OOD inputs (Hendrycks et al., 2019b; Mohseni et al., 2020; Sehwag et al., 2021).

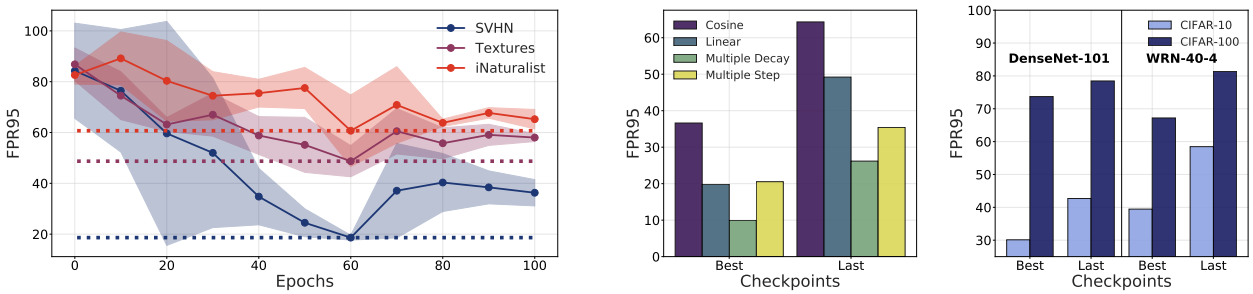
Despite the promising results achieved by previous methods (Hendrycks & Gimpel, 2017; Hendrycks et al., 2019b; Liu et al., 2020; Ming et al., 2022b), limited attention is paid to considering whether the given well-trained model is the most appropriate basis for OOD detection. In general, models deployed for various applications have different original targets (e.g., multi-class classification (Goodfellow et al., 2016)) instead of OOD detection (Nguyen et al., 2015). However, most representative score functions, e.g., MSP (Hendrycks et al., 2019b), ODIN (Liang et al., 2018), and Energy (Liu et al., 2020), uniformly leverage the given models for OOD detection (Yang et al., 2021). The above target-oriented discrepancy naturally motivates the following critical question: *does the given well-trained model have the optimal OOD discriminative capability?* If not, *how can we find a more appropriate counterpart for OOD detection?*

In this work, we start by revealing an interesting empirical observation, i.e., there always exists a historical training stage where the model has a higher OOD detection performance than the final well-trained one (as shown in Figure 1), spanning among different OOD/ID datasets (Netzer et al., 2011; Van Horn et al., 2018) under different learning rate schedules (Loshchilov & Hutter, 2017) and model structures (Huang et al., 2017; Zagoruyko & Komodakis, 2016). It shows the inconsistency between gaining better OOD discriminative capability (Nguyen et al., 2015) and pursuing better performance on ID data during training. Through the in-depth analysis from various perspectives (as illustrated in Figure 2), we figure out one possible attribution at the data level is memorizing the *atypical samples* (compared with others at the semantic level) that are hard to generalize for the model. Seeking zero training error on those samples leads the model more confident in the unseen OOD inputs.

The above analysis inspires us to propose a new method, namely, *Unleashing Mask* (UM), to excavate the overlaid detection capability of a well-trained given model by al-

[1]Department of Computer Science, Hong Kong Baptist University [2]CMIC, Shanghai Jiao Tong University [3]Shanghai AI Laboratory [4]Mohamed bin Zayed University of Artificial Intelligence [5]Sydney AI Centre, The University of Sydney. Correspondence to: Bo Han <bhanml@comp.hkbu.edu.hk>, Jiangchao Yao <Sunarker@sjtu.edu.cn>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).



(a) Curves of FPR95 based on Energy score (b) Diff. LR Schedules (c) Diff. Model Structures

Figure 1. **Critical reveal of the intermediate stage with better OOD detection performance across various setups:** (a) the curves of FPR95 (false positive rate of OOD examples when the true positive rate of ID examples is at 95%) based on Energy score (Liu et al., 2020) across three different OOD datasets during the training on CIFAR-10 dataset; (b) comparison of best/last checkpoints for OOD detection under different lr schedules on CIFAR-10; (c) comparison of best/last checkpoints for OOD detection under different model structures on CIFAR-10/CIFAR-100. The results are obtained after multiple runs, and we leave other setup details to Section 4.1 and Appendix G.1.

leviating the memorization of those atypical samples (as illustrated in Figure 3) of ID data. In general, we aim to backtrack its previous stage with better OOD discriminative capabilities. To achieve this target, there are two essential issues: (1) *the model that is well-trained on ID data has already memorized some atypical samples*; (2) *how to forget those memorized atypical samples considering the given model?* Accordingly, our proposed UM contains two parts utilizing different insights to address the two problems. First, as atypical samples are more sensitive to the change of model parameters, we initialize a mask with the specific cutting rate to mine these samples with constructed parameter discrepancy. Second, with the loss reference estimated by the mask, we conduct the constrained gradient ascent for model forgetting (i.e., Eq. (3)). It will encourage the model to finally stabilize around the optimal stage. To avoid severe sacrifices of the original task performance on ID data, we further propose *UM Adopts Pruning* (UMAP) which tunes on the introduced mask with the newly designed objective.

We conduct extensive experiments (in Section 4 and Appendixes G.1 to G.10) to present the working mechanism of our proposed methods. We have verified the effectiveness with a series of OOD detection benchmarks mainly on two common ID datasets, i.e., CIFAR-10 and CIFAR-100. Under the various evaluations, our UM, as well as UMAP, can indeed excavate the better OOD discriminative capability of the well-trained given models and the averaged FPR95 can be reduced by a significant margin. Finally, a range of ablation studies, verification on the ImageNet pretrained model, and further discussions from both empirical and theoretical views are provided. Our main contributions are as follows,

- Conceptually, we explore the OOD detection performance via a new perspective, i.e., backtracking the initial model training phase without regularizing by any auxiliary outliers, different from most previous works that start with the well-trained model on ID data.
- Empirically, we reveal the potential OOD discriminative capability of the well-trained model, and figure out one data-level attribution of concealing it during original training is memorizing the atypical samples.
- Technically, we propose a novel *Unleashing Mask* (UM) and its practical variant UMAP, which utilizes the newly designed forgetting objective with ID data to excavate the intrinsic OOD detection capability.
- Experimentally, we conduct extensive explorations to verify the overall effectiveness of our method in improving OOD detection performance, and perform various ablations to provide a thorough understanding.

2. Preliminaries

We consider multi-class classification as the original training task (Nguyen et al., 2015), where $\mathcal{X} \subset \mathbb{R}^d$ denotes the input space and $\mathcal{Y} = \{1, \dots, C\}$ denotes the label space. In practical, a reliable classifier should be able to figure out the OOD input, which can be considered as a binary classification problem. Given $\mathcal{P}$, the distribution over $\mathcal{X} \times \mathcal{Y}$, we consider $\mathcal{D}_{\text{in}}$ as the marginal distribution of $\mathcal{P}$ for $\mathcal{X}$, namely, the distribution of ID data. At test time, the environment can present a distribution $\mathcal{D}_{\text{out}}$ over $\mathcal{X}$ of OOD data. In general, the OOD distribution $\mathcal{D}_{\text{out}}$ is defined as an irrelevant distribution of which the label set has no intersection with $\mathcal{Y}$ (Yang et al., 2021) and thus should not be predicted by the model. A decision can be made with the threshold λ:

$$D_\lambda(x; f) = \begin{cases} \text{ID} & S(x) \geq \lambda \\ \text{OOD} & S(x) < \lambda \end{cases}, \tag{1}$$

Building upon the model $f \in \mathcal{H} : \mathcal{X} \to \mathbb{R}^c$ trained on ID data with the logit outputs, the goal of decision is to utilize the scoring function $S : \mathcal{X} \to \mathbb{R}$ to distinguish the inputs of $\mathcal{D}_{\text{in}}$ from that of $\mathcal{D}_{\text{out}}$ by $S(x)$. Typically, if the score value is larger than the threshold λ, the associated input

x is classified as ID and vice versa. We consider several representative scoring functions designed for OOD detection, e.g., MSP (Hendrycks & Gimpel, 2017), ODIN (Liang et al., 2018), and Energy (Liu et al., 2020). More detailed definitions and implementation are provided in Appendix A.

To mitigate the issue of over-confident predictions for some OOD data (Hendrycks & Gimpel, 2017; Liu et al., 2020), recent works (Hendrycks et al., 2019b; Tack et al., 2020) utilize the auxiliary unlabeled dataset to regularize the model behavior. Among them, one representative baseline is outlier exposure (OE) (Hendrycks et al., 2019b). OE can further improve the detection performance by making the model $f(\cdot)$ finetuned from a surrogate OOD distribution $\mathcal{D}^{\text{s}}_{\text{out}}$, and its corresponding learning objective is defined as follows,

$$\mathcal{L}_f = \mathbb{E}_{\mathcal{D}_{\text{in}}}\left[\ell_{\text{CE}}(f(x), y)\right] + \lambda \mathbb{E}_{\mathcal{D}^{\text{s}}_{\text{out}}}\left[\ell_{\text{OE}}(f(x))\right], \quad (2)$$

where λ is the balancing parameter, $\ell_{\text{CE}}(\cdot)$ is the Cross-Entropy (CE) loss, and $\ell_{\text{OE}}(\cdot)$ is the Kullback-Leibler divergence to the uniform distribution, which can be written as $\ell_{\text{OE}}(h(\boldsymbol{x})) = -\sum_k \texttt{softmax}_k\ f(x)/C$, where $\texttt{softmax}_k(\cdot)$ denotes the k-th element of a softmax output. The OE loss $\ell_{\text{OE}}(\cdot)$ is designed for model regularization, making the model learn from surrogate OOD inputs to return low-confident predictions (Hendrycks et al., 2019b).

Although previous works show promising results via designing scoring functions or regularizing models with different auxiliary outlier data, few of them investigated or excavated the original discriminative capability of the well-trained model using ID data. In this work, we introduce the layer-wise mask m (Han et al., 2016; Ramanujan et al., 2020) to mine the atypical samples that are memorized by the model. Accordingly, the decision can be rewritten as $D(x; m \odot f)$, and the output of a masked model is defined as $m \odot f(x)$.

3. Proposed Method: Unleashing Mask

In this section, we introduce our new method, i.e., *Unleashing Mask* (UM), to reveal the potential OOD discriminative capability of the well-trained model. First, we present and discuss the important observation that inspires our methods (Section 3.1). Second, we provide the insights behind the two critical parts of our UM (Section 3.2). Lastly, we introduce the overall framework and its learning objective, as well as a practical variant of UM, i.e., UMAP (Section 3.3).

3.1. Overlaid OOD Detection Capability

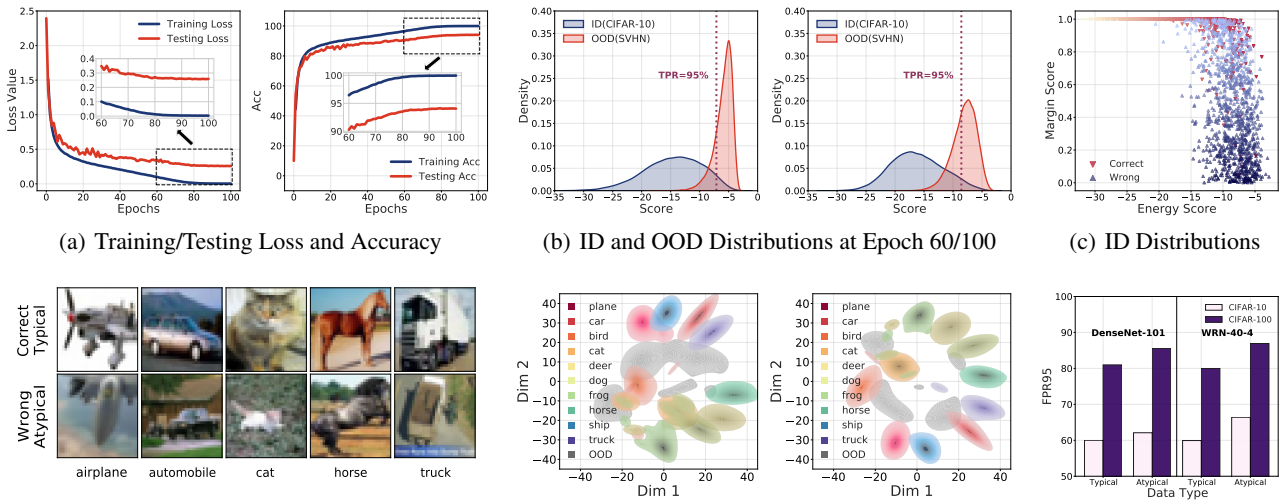
First, we present the phenomenon of the inconsistency between pursuing better OOD discriminative capability and smaller training errors during the original task. Empirically, as shown in Figure 1, we trace the OOD detection performance during the model training after multiple runs of the experiments. Across three different OOD datasets in Figure 1(a), we can observe the existence of a better detection performance using the index of FPR95 metric based on the Energy (Liu et al., 2020) score. The generality has also been demonstrated under different learning schedules, model structures, and ID datasets in Figures 1(b) and 1(c). Without any auxiliary outliers, it motivates us to explore the underlying mechanism of the training process with ID data.

We further delve into the learning dynamics from various perspectives in Figure 2, and we reveal the critical data-level attribution for the OOD discriminative capability. In Figure 2(a), we find that the training loss has reached a reasonably small value[1] at Epoch 60 where its detection performance achieves a satisfactory level. However, if we further minimize the training loss, the trend of the FPR95 curve shows almost the opposite direction with both training and testing loss or accuracy (see Figures 1(a) and 2(a)). The comparison of the ID/OOD distributions is presented in Figure 2(b). To be specific, the statics of the two distributions indicate that the gap between the ID and OOD data gets narrow as their overlap grows along with the training. After Epoch 60, although the model becomes more confident on ID data which satisfies a part of the calibration target (Hendrycks et al., 2019a), its predictions on the OOD data also become more confident which is unexpected. Using the margin value defined in logit space (see Eq. (14)), we gather the statical with Energy score in Figure 2(c). The misclassified samples are found to be close to the decision boundary and have a high uncertainty level in model prediction. Accordingly, we extract those samples that were learned by the model at this period. As shown in Figures 2(d), 2(e) and 2(f), the misclassified samples learned after Epoch 60 present much atypical semantic features, which results in more diverse feature embedding and may impair OOD detection. As deep neural networks tend to first learn the data with typical features (Arpit et al., 2017), we attribute the inconsistent trend to memorizing those atypical data at the later stage.

3.2. Unleashing the Potential Discriminative Power

In general, the models that are developed for the original classification tasks are always seeking better performance (e.g., higher testing accuracy and lower training loss) in practice. However, the inconsistent trend revealed before provides us the possibility to unleash the potential detection power only considering the ID data in training. To this end, we have two important issues that need to address: (1) *the well-trained model may have already memorized some atypical samples which cannot be figured out*; (2) *how to forget those atypical samples considering the given model?*

[1]Note that it is not the conventional overfitting (Goodfellow et al., 2016) as the testing loss is still decreasing. In Section 4.3 and Appendix D, we provide both the empirical comparison of some targeted strategies and the conceptual comparison of them.



(a) Training/Testing Loss and Accuracy (b) ID and OOD Distributions at Epoch 60/100 (c) ID Distributions

(d) Wrongly/Correctly Classified Data at Epoch 60 (e) TSNE Visualization at Epoch 60/100 (f) Diff. Effects

Figure 2. **Delve into the data-level attribution of the phenomenon with the original multi-classification on CIFAR-10:** (a) training/testing loss and accuracy on ID data; (b) comparison of ID/OOD distributions based on Energy score at Epoch 60/100 (c) scatter plot of wrongly/correctly classified samples at Epoch 60 using Margin value and Energy score (d) visualization of wrongly/correctly classified samples at Epoch 60; (e) TSNE visualization of the feature embedding on ID/OOD data at Epoch 60/100. (f) Effects on OOD detection of tuning with those identified typical/atypical samples, more detailed setup, and results can be referred to in Appendix G.2. Through comparison from various perspectives, we find that achieving a reasonably small loss value (at round Epoch 60) on ID data is enough for OOD detection. However, continually optimizing on those atypical samples (e.g., wrongly classified in (d)) may impair OOD detection.

Atypical mining with constructed discrepancy. As shown Figures 2(a) and 2(b), the training statics provide limited information to accurately differentiate the stage that learns on typical or atypical data. We thus explore to construct the parameter discrepancy to mine the atypical samples from a well-trained given model in the light of the learning dynamics (Goodfellow et al., 2016; Arpit et al., 2017) of deep neural networks and the model uncertainty representation (Gal & Ghahramani, 2016). Specifically, we employ a randomly initialized layer-wise mask which applied to all layers. It is consistent with the mask generation in the conventional pruning pipeline (Han et al., 2016). In Figure 3, we provide empirical evidence to show that we can figure out atypical samples by a certain mask ratio δ, through which we can gradually mine the model stage that misclassifies atypical samples. We provide more discussion about the underlying intuition of masking in Appendix E.

Model forgetting with gradient ascent. As the training loss achieves zero at the final stage of the given model, we need extra optimization signals to forget those memorized atypical samples. Considering the previous consistent trend before the potential optimal stage (e.g., before Epoch 60 in Figure 1(a)), the optimization signal also needs to control the model update not to be too greedy to drop the discriminative features that can be utilized for OOD detection. Starting with the well-trained given model, we can employ the gradient ascent (Sorg et al., 2010; Ishida et al., 2020) to forget the targeted samples, while the tuning phase should also prevent further updates if it achieves the expected stage. As for another implementation choice, e.g., retraining the model from scratch for our targets, we discuss it in Appendix G.3.

3.3. Method Realization

Based on previous insights, we present our overall framework and the learning objective of the proposed UM and UMAP for OOD detection. Lastly, we discuss their compatibility with either the fundamental scoring functions or the outlier exposure approaches utilizing auxiliary outliers.

Framework. As illustrated in Figure 3, our framework consists of two critical components for uncovering the intrinsic OOD detection capability: (1) the initialized mask with a specific masking rate for constructing the output discrepancy with the original model; (2) the subsequent adjustment for alleviating the memorization of atypical samples. The overall workflow starts with estimating the loss value of misclassifying those atypical samples and then conducts tuning on the model or the masked output to forget them.

Forgetting via Unleashing Mask (UM). Based on previous insights, we introduce the forgetting objective as,

$$\min \mathcal{L}_{\text{UM}} = \min_{m_\delta \in [0,1]^n} |\ell_{\text{CE}}(f) - \widehat{\ell}_{\text{CE}}(m_\delta \odot f^*)| + \widehat{\ell}_{\text{CE}}(m_\delta \odot f^*), \tag{3}$$

where m_δ is the layer-wise mask with the masking rate δ, ℓ_{CE} is the CE loss, $\widehat{\ell}_{\text{CE}}$ is the averaged CE loss over the ID

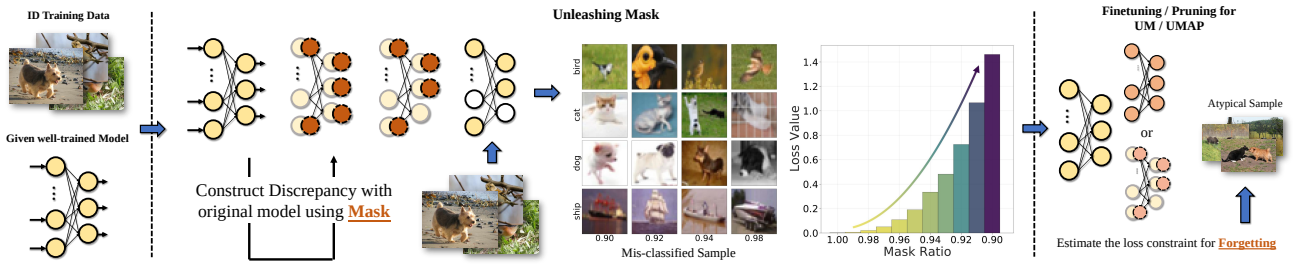


Figure 3. **Overview of Unleashing Mask:** Given a well-trained model, we initialize a mask for mining the atypical samples that are sensitive to the changes in model parameters. Then we finetune the original model or adopt pruning with the estimated forgetting threshold, i.e., the loss value estimated by the UM. The final model can serve as the base of those representative score functions to utilize the discriminative features and also as a new initialization of outlier exposure methods. In addition, we also present examples of misclassified samples in ID data after masking the original well-trained model, and loss value using the masked outputs w.r.t. different mask ratios.

training data, $|\cdot|$ indicates the computation for absolute value and $m_\delta \odot f^*$ denotes the masked output of the fixed pretrained model that is used to estimate the loss constraint for the learning objective of forgetting. The value of $\widehat{\ell}_{\text{CE}}(\cdot)$ would be constant during the whole finetuning process. Concretely, the well-trained model will start to optimize itself again if it memorizes the atypical samples and achieves almost zero loss value. We provide a positive gradient signal when the current loss value is lower than the estimated one and vice versa. The model is expected to finally stabilize around the stage that can forget those atypical samples. To be more specific, for a mini-batch of ID samples, they are forwarded to the (pre-trained) model and the loss is automatically computed in Eq. (3). Based on our introduced layer-wise mask, the atypical samples would be easier to induce large loss values than the rest, and will be forced to be wrongly classified in the end-to-end optimization, in which atypical samples are forgotten without being identified.

Unleashing Mask Adopts Pruning (UMAP). Considering the potential negative effect on the original task performance when conducting tuning for forgetting, we further propose a variant of UM Adopts Pruning, i.e., UMAP, to conduct tuning based on the masked output (e.g., replace $\ell_{\text{CE}}(f)$ to $\ell_{\text{CE}}(\hat{m}_p \odot f)$ in Eq 3) using a functionally different mask $\hat{m}_p$ with its pruning rate p as follows,

$$\min \mathcal{L}_{\text{UMAP}} = \min_{\substack{\hat{m}_p \in [0,1]^n \\ m_\delta \in [0,1]^n}} |\ell_{\text{CE}}(\hat{m}_p \odot f) - \widehat{\ell}_{\text{CE}}(m_\delta \odot f^*)| + \widehat{\ell}_{\text{CE}}(m_\delta \odot f^*), \tag{4}$$

Different from the objective of UM (i.e., Eq 3) that minimizes the loss value over the model parameter, the objective of UMAP minimizes the loss over the mask $\hat{m}_p$ to achieve the target of forgetting atypical samples. UMAP provides an extra mask to restore the detection capacity but doesn't affect the model parameter for the inference on original tasks, indicating that UMAP is a more practical choice in real-world applications (as empirically verified in our experiments like Table 1). We present the algorithms of UM (in Algorithm 1) and UMAP (in Algorithm 2) in Appendix F.

Compatible with other methods. As we explore the original OOD detection capability of the well-trained model, it is orthogonal and compatible with those promising methods that equip the given model with better detection ability. To be specific, through our proposed methods, we reveal the overlaid OOD detection capability by tuning the original model toward its intermediate training stage. The discriminative feature learned at that stage can be utilized by different scoring functions (Huang et al., 2021; Liu et al., 2020; Sun & Li, 2022), like ODIN (Liang et al., 2018) adopted in Figure 4(c). For those methods (Hendrycks et al., 2019a; Liu et al., 2020; Ming et al., 2022b) utilizing the auxiliary outliers to regularize the model, our finetuned model obtained by UM and UMAP can also serve as their starting point or adjustment. As our method does not require any auxiliary outlier data to be involved in training, adjusting the model using ID data during its developing phase is practical.

4. Experiments

In this section, we present the performance comparison of the proposed method in the OOD detection scenario. Specifically, we verify the effectiveness of our UM and UMAP with two mainstreams of OOD detection approaches: (i) fundamental scoring function methods; (ii) outlier exposure methods involving auxiliary samples. To better understand our proposed method, we further conduct various explorations on the ablation study and provide the corresponding discussion on each sub-aspect considered in our work. More details and additional results are presented in Appendix G.

4.1. Experimental Setups

Datasets. Following the common benchmarks used in previous work (Liu et al., 2020; Ming et al., 2022b), we adopt

Table 1. Main Results (%). Comparison with competitive OOD detection baselines. (averaged by multiple trials)

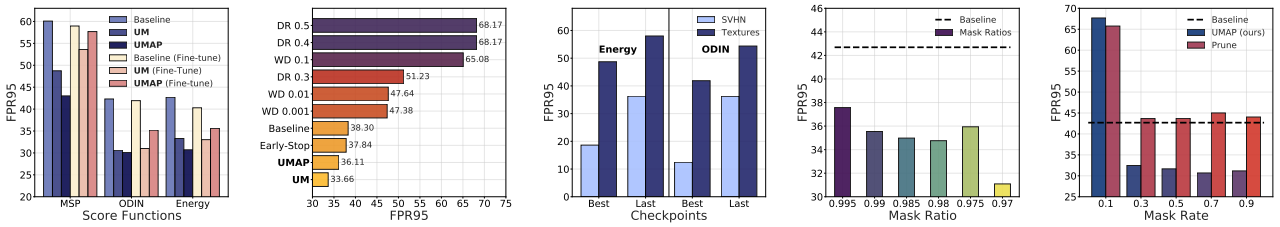
$\mathcal{D}_{\text{in}}$	Method	AUROC↑	AUPR↑	FPR95↓	ID-ACC↑	w./w.o $\mathcal{D}_{\text{aux}}$
CIFAR-10	MSP(Hendrycks & Gimpel, 2017)	89.90 ± 0.30	91.48 ± 0.43	60.08 ± 0.76	$\mathbf{94.01 \pm 0.08}$	
	ODIN(Liang et al., 2018)	91.46 ± 0.56	91.67 ± 0.58	42.31 ± 1.38	$\mathbf{94.01 \pm 0.08}$	
	Mahalanobis(Lee et al., 2018b)	75.10 ± 1.04	72.32 ± 1.92	61.35 ± 1.25	$\mathbf{94.01 \pm 0.08}$	
	Energy(Liu et al., 2020)	92.07 ± 0.22	92.72 ± 0.39	42.69 ± 1.31	$\mathbf{94.01 \pm 0.08}$	
	Energy+UM (ours)	93.73 ± 0.36	94.27 ± 0.60	33.29 ± 1.70	92.80 ± 0.47	
	Energy+UMAP (ours)	$\mathbf{93.97 \pm 0.11}$	$\mathbf{94.38 \pm 0.06}$	$\mathbf{30.71 \pm 1.94}$	$\mathbf{94.01 \pm 0.08}$	
	OE(Hendrycks et al., 2019b)	97.07 ± 0.01	97.31 ± 0.05	13.80 ± 0.28	92.59 ± 0.32	✓
	Energy (w. $\mathcal{D}_{\text{aux}}$)(Liu et al., 2020)	94.58 ± 0.64	94.69 ± 0.65	18.79 ± 2.31	80.91 ± 3.13	✓
	POEM(Ming et al., 2022b)	94.37 ± 0.07	94.51 ± 0.06	18.50 ± 0.33	77.24 ± 2.22	✓
	OE+UM (ours)	$\mathbf{97.60 \pm 0.03}$	$\mathbf{97.87 \pm 0.02}$	$\mathbf{11.22 \pm 0.16}$	93.66 ± 0.12	✓
	OE+UMAP (ours)	97.48 ± 0.01	97.74 ± 0.00	12.21 ± 0.09	$\mathbf{94.01 \pm 0.08}$	✓
CIFAR-100	MSP(Hendrycks & Gimpel, 2017)	74.06 ± 0.69	75.37 ± 0.73	83.14 ± 0.87	$\mathbf{74.86 \pm 0.21}$	
	ODIN(Liang et al., 2018)	76.18 ± 0.14	76.49 ± 0.20	78.93 ± 0.31	$\mathbf{74.86 \pm 0.21}$	
	Mahalanobis(Lee et al., 2018b)	63.90 ± 1.91	64.31 ± 0.91	78.79 ± 0.50	$\mathbf{74.86 \pm 0.21}$	
	Energy(Liu et al., 2020)	$\mathbf{76.29 \pm 0.24}$	$\mathbf{77.06 \pm 0.55}$	78.46 ± 0.06	$\mathbf{74.86 \pm 0.21}$	
	Energy+UM (ours)	76.22 ± 0.42	76.39 ± 1.03	74.05 ± 0.55	64.55 ± 0.24	
	Energy+UMAP (ours)	75.57 ± 0.59	75.66 ± 0.07	$\mathbf{72.21 \pm 1.46}$	$\mathbf{74.86 \pm 0.21}$	
	OE(Hendrycks et al., 2019b)	90.55 ± 0.87	90.34 ± 0.94	34.73 ± 3.85	73.59 ± 0.30	✓
	Energy (w. $\mathcal{D}_{\text{aux}}$)(Liu et al., 2020)	88.92 ± 0.57	89.13 ± 0.56	37.90 ± 2.59	57.85 ± 2.65	✓
	POEM(Ming et al., 2022b)	88.95 ± 0.54	88.94 ± 0.31	38.10 ± 1.30	56.18 ± 1.92	✓
	OE+UM (ours)	91.04 ± 0.11	$\mathbf{91.13 \pm 0.24}$	34.71 ± 0.81	$\mathbf{75.15 \pm 0.18}$	✓
	OE+UMAP (ours)	$\mathbf{91.10 \pm 0.16}$	90.99 ± 0.23	$\mathbf{33.62 \pm 0.26}$	74.76 ± 0.11	✓

Table 2. Fine-grained Results (%). Comparison on different OOD benchmark datasets. (averaged by multiple trials)

ID dataset	**Method**	**OOD dataset**					
		CIFAR-100		**Textures**		**Places365**	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CIFAR-10	MSP	66.43 ± 1.25	87.73 ± 0.02	65.20 ± 1.33	88.06 ± 0.61	61.34 ± 0.60	89.63 ± 0.15
	ODIN	55.31 ± 0.85	87.75 ± 0.37	53.11 ± 4.84	87.13 ± 2.04	43.77 ± 0.20	91.70 ± 0.30
	Mahalanobis	81.61 ± 0.96	64.52 ± 0.73	$\mathbf{20.04 \pm 1.43}$	$\mathbf{94.38 \pm 0.78}$	86.21 ± 1.36	64.00 ± 1.21
	Energy	54.65 ± 1.24	$\mathbf{89.01 \pm 1.18}$	57.09 ± 3.52	87.51 ± 1.43	38.62 ± 1.64	93.03 ± 0.20
	Energy+UM (ours)	$\mathbf{54.62 \pm 1.16}$	88.30 ± 0.30	41.61 ± 3.67	91.31 ± 0.01	$\mathbf{30.85 \pm 0.58}$	$\mathbf{94.27 \pm 0.16}$
	Method	**SUN**		**LSUN**		**iNaturalist**	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
	MSP	60.27 ± 0.66	90.00 ± 0.24	36.43 ± 1.94	95.17 ± 0.32	67.53 ± 1.64	88.01 ± 0.82
	ODIN	41.14 ± 1.29	92.34 ± 0.62	5.16 ± 0.76	98.96 ± 0.09	54.41 ± 0.91	90.17 ± 0.19
	Mahalanobis	84.56 ± 1.51	66.41 ± 4.57	69.18 ± 3.52	66.41 ± 4.57	80.76 ± 2.48	71.77 ± 1.12
	Energy	36.73 ± 1.72	93.63 ± 0.34	6.25 ± 0.43	98.77 ± 0.07	59.11 ± 1.18	89.71 ± 0.06
	Energy+UM (ours)	$\mathbf{27.88 \pm 0.73}$	$\mathbf{94.83 \pm 0.11}$	$\mathbf{2.91 \pm 0.53}$	$\mathbf{99.22 \pm 0.11}$	$\mathbf{46.27 \pm 2.74}$	$\mathbf{92.75 \pm 0.80}$

`CIFAR-10`, `CIFAR-100` (Krizhevsky, 2009) as our major ID datasets, and we also adopt `ImageNet` (Deng et al., 2009) for performance exploration. We use a series of different image datasets as the OOD datasets, e.g., `Textures` (Cimpoi et al., 2014), `Places365` (Zhou et al., 2017), `SUN` (Xiao et al., 2010), `LSUN` (Yu et al., 2015), `iNaturalist` (Van Horn et al., 2018) and `SVHN` (Netzer et al., 2011). We also use the other ID dataset as OOD dataset when training on a specific ID dataset, given that none of them shares the same classes, e.g., we treat `CIFAR-100` as the OOD dataset when training on `CIFAR-10` for comparison. We utilize the ImageNet-1k (Deng et al., 2009) training set as the auxiliary dataset for all of our experiments about fine-tuning with auxiliary outliers (e.g., OE/Energy/POEM), which is detailed in Appendix G.1. This choice follows previous literature (Hendrycks et al., 2019b; Liu et al., 2020; Ming et al., 2022b) that considers the dataset's availability and the absence of any overlap with the ID datasets.

Evaluation metrics. We employ the following three common metrics to evaluate the performance of OOD detection: (i) Area Under the Receiver Operating Characteristic curve (AUROC) (Davis & Goadrich, 2006) can be interpreted as the probability for a positive sample to have a higher discriminating score than a negative sample (Fawcett, 2006); (ii) Area Under the Precision-Recall curve (AUPR) (Manning & Schütze, 1999) is an ideal metric to adjust the extreme difference between positive and negative base rates; (iii)



(a) Train v.s. finetune (b) Compare with Overfitting (c) Diff. Score Functions (d) Diff. Masking Ratios (e) UMAP v.s. Prune

Figure 4. **Ablation studies:** (a) efficiency of the finetuning adopted in UM and UMAP; (b) comparison of UM and UMAP with other strategies for alleviating the conventional overfitting; (c) the historical model stage using different scoring functions for OOD detection; (d) effects of using different masking ratios for atypical mining in UM; (e) comparison of using vanilla pruning with our proposed UMAP.

False Positive Rate (FPR) at 95% True Positive Rate (TPR) (Liang et al., 2018) indicates the probability for a negative sample to be misclassified as positive when the true positive rate is at 95%. We also include in-distribution testing accuracy (ID-ACC) to reflect the preservation level of the performance for the original classification task on ID data.

OOD detection baselines. We compare the proposed method with several competitive baselines in the two directions. Specifically, we adopt Maximum Softmax Probability (MSP) (Hendrycks & Gimpel, 2017), ODIN (Liang et al., 2018), Mahalanobis score (Lee et al., 2018b), and Energy score (Liu et al., 2020) as scoring function baselines; We adopt OE (Hendrycks et al., 2019b), Energy-bounded learning (Liu et al., 2020), and POEM (Ming et al., 2022b) as baselines with outliers. For all scoring function methods, we assume the accessibility of well-trained models. For all methods involving outliers, we constrain all major experiments to a finetuning scenario, which is more practical in real cases. Different from training a dual-task model at the very beginning, equipping deployed models with OOD detection ability is a much more common circumstance, considering the millions of existing deep learning systems. We leave more implementation details in Appendix A.

4.2. Performance Comparison

In this part, we present the performance comparison with some representative baseline methods to demonstrate the effectiveness of our UM and UMAP. In each category of Table 1, we choose one with the best detection performance to adopt UM or UMAP and check the three evaluation metrics of OOD detection and the ID-ACC.

In Table 1, we summarize the results using different methods. For the scoring-based methods, our UM can further improve the overall detection performance by alleviating the memorization of atypical ID data, when the ID-ACC keeps comparable with the baseline. For the complex CIFAR-100 dataset, our UMAP can be adopted as a practical way to empower the detection performance and simultaneously avoid severely affecting the original performance on ID data. As for those methods of the second category (i.e., involving auxiliary outlier $\mathcal{D}_{\text{aux}}$ sampled from ImageNet), since we consider a practical workflow, i.e., fine-tuning, on the given model, OE achieves the best performance on the task. Due to the special optimization characteristic, Energy (w. $\mathcal{D}_{\text{aux}}$) and POEM focus more on the energy loss on differentiating OOD data while performing not well on the preservation of ID-ACC. Without sacrificing much performance on ID data, OE with our UM can still achieve better detection performance. In Table 2, the fine-grained detection performance on each OOD testing set demonstrates the general effectiveness of UM and UMAP. Note that we may observe Mahalanobis can sometimes achieve the best performance on the specific OOD test set (e.g., Textures). It is probably because Mahalanobis is prone to overfitting on texture features during fine-tuning with Textures. In contrast, according to Table 2, Mahalanobis achieves the worst results on the other five datasets. We leave more results (e.g., completed comparison in Table 10; more fine-grained results in Tables 19 and 20; using another model structure in Tables 21, 22 and 23) to Appendix, which has verified the significant improvement (up to 18% reduced on averaged FPR95) across various setups and also on a large-scale ID dataset (i.e., ImageNet (Deng et al., 2009) in Table 11).

4.3. Ablation and Further Analysis

In this part, we conduct further explorations and analysis to provide a thorough understanding of our UM and UMAP. Moreover, we also provide additional experimental results about further explorations on OOD detection in Appendix G.

Practicality of the considered setting and the implementation choice. Following the previous work (Liu et al., 2020; Hendrycks et al., 2019b), we consider the same setting that starts from a given well-trained model in major explorations, which is practical but can be extended to another implementation choice, i.e., retraining the whole model. In Figure 4(a), we show the effectiveness of UM/UMAP under different choices. It is worth noting that UM adopting fine-

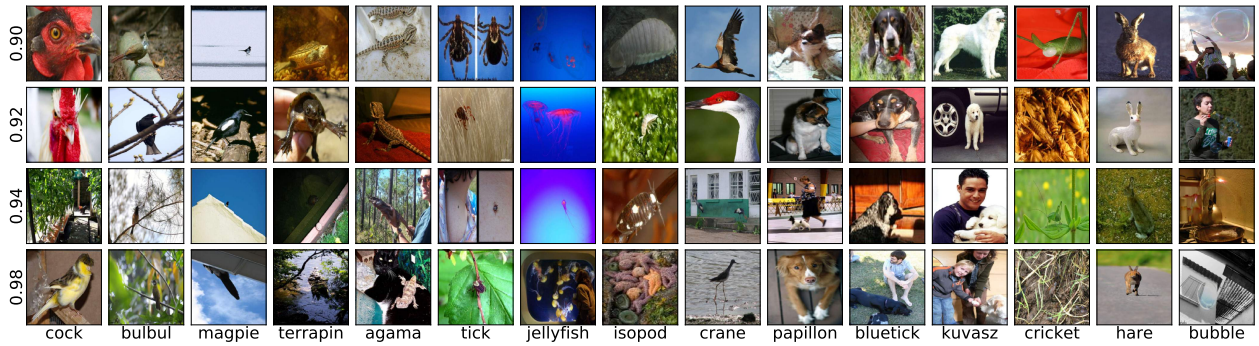


Figure 5. **Sample Visualizations:** examples of the misclassified samples after adopting masking on the original well-trained model using the ImageNet dataset. At the left of each line, we indicate the mask ratio that is adopted on the model. We can find that masking with a smaller ratio forces the model to misclassify simple samples (clear contours around subjects, single color background) while masking with a larger ratio guides the model to misclassify complex samples (unclear contours, noisy background). This inspection empirically verifies our intuition using the proper mask ratio to identify those atypical samples and then force the model to forget them.

tuning has shown the advantages of being cost-effective on convergence compared with train-from-scratch, which we leave more discussion and comparison in Appendix G.3.

Specificity and applicability of excavated OOD discriminative capability. As mentioned before, the intrinsic OOD discriminative capability is distinguishable from conventional overfitting. We empirically compare UM/UMAP with dropout (DR), weight decay (WD), and early stop in Figure 4(b). UM gain lower FPR95 from the newly designed objective for forgetting. In Figure 4(c), we present the applicability of the OOD detection capability using different score functions, which implies the generated model stage better meets the requirement of uncertainty estimation.

Effects of the mask on mining atypical samples. In Figure 4(d), we compare UM with different mask ratios for mining the atypical samples, which seeks to find the intermediate model stage that wrongly classified the atypical samples. The results show reasonably small ratios (e.g., from 0.995 to 0.97) that we knocked off in the original model can help us to achieve the targets. More detailed analysis of the mask ratio and the discussion about the underlying intuition of atypical mining are provided in Appendixes G.10 and E.

Exploration on UMAP and vanilla model pruning. Although the large constraint on training loss can help reveal the OOD detection performance, the ID-ACC may be undermined under such circumstances. To mitigate this issue, we further adopt pruning in UMAP to learn a mask instead of tuning the model parameters directly. In Figure 4(e), we explore various prune rates p and demonstrate their effectiveness. Specifically, our UMAP can achieve a lower FPR95 than vanilla pruning with the original objective. The prune rate can be selected from a wide range (e.g., $p \in [0.3, 0.9]$) to guarantee a fast convergence and effectiveness. We also

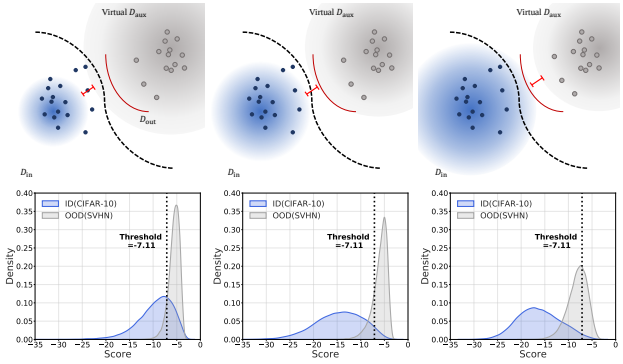


Figure 6. Illustration about the framework for theoretical anslysis. From left to right: the model under-represent on $\mathcal{D}_{\text{in}}$ with the lower confidence on the atypical samples close to the boundary; the model reaches a near-optimal representation status on $\mathcal{D}_{\text{in}}$; the model over-represent on $\mathcal{D}_{\text{in}}$. See Figure 7 for more explanations.

provide additional discussion on UMAP in Appendix G.9.

Sample visualization of the atypical samples identified by our mask. In Figure 5, we visualize the misclassified samples using the ImageNet (Deng et al., 2009) dataset with the pre-trained model by adopting different mask ratios. We can find that masking the model constructs the parameter discrepancy, which helps us to identify some ID samples with atypical semantic information (e.g., those samples in the bottom line compared with the above in each class). It demonstrates the rationality of our intuition to adopt masking. We leave more visualization results in Appendix E.

Theoretical insights on ID data property. Similar to prior works (Lee et al., 2018a; Sehwag et al., 2021), here we present the major results based on the sample complexity analysis adopted in POEM (Ming et al., 2022b). Due to the limited space, please refer to Appendix B for the completed

analysis and Figure 7 for more conceptual understanding.

Theorem 4.1. *Given a simple Gaussian mixture model in binary classification with the hypothesis class $\mathcal{H} = sign(\theta^T x), \theta \in \mathbb{R}^d$. There exists constant α, δ^* and ϵ that,*

$$\frac{\mu^T \theta^*_{n_1,n_2}}{\sigma ||\theta^*_{n_1,n_2}||} \geq \frac{||\mu||^2 - \sigma^{\frac{1}{2}} ||\mu||^{\frac{3}{2}} - \frac{\sigma^2(|\alpha-\delta^*|+\epsilon)}{2}}{2\sqrt{\frac{\sigma^2}{n}(d+\frac{1}{\sigma}) + ||\mu||^2}} \tag{5}$$

Since $\text{FPR}(\theta^*_{n_1,n_2}) = erf(\frac{\mu^T \theta^*_{n_1,n_2}}{\sigma ||\theta^*_{n_1,n_2}||})$ is monotonically decreasing, as the lower bound of $\frac{\mu^T \theta^*_{n_1,n_2}}{\sigma ||\theta^*_{n_1,n_2}||}$ will increase with the constraint of $|\alpha - \delta^*|$ (which corresponds to the illustrated distance from the outlier boundary in right-most of Figure 6) decrease in our methods, the upper bound of $\text{FPR}(\theta^*_{n_1,n_2})$ will decrease. One insight is learning more atypical ID samples needs more high-quality auxiliary outliers (near ID data) to shape the OOD detection capability.

Additional experimental results of explorations. Except for the major performance comparisons and the previous ablations, we also provide further discussion and analysis from different views in Appendix G, including the practicality of the considered setting, the effects of the mask on mining atypical samples, discussion of UMAP with vanilla pruning, additional comparisons with more advanced methods and completed results of our proposed UM and UMAP.

5. Related Work

OOD Detection without auxiliary data. Hendrycks & Gimpel (2017) formally shed light on out-of-distribution detection, proposing to use softmax prediction probability as a baseline which is demonstrated to be unsuitable for OOD detection (Hendrycks et al., 2019b). Subsequent works (Sun et al., 2021) keep focusing on designing post-hoc metrics to distinguish ID samples from OOD samples, among which ODIN (Liang et al., 2018) introduces small perturbations into input images to facilitate the separation of softmax score, Mahalanobis distance-based confidence score (Lee et al., 2018b) exploits the feature space by obtaining conditional Gaussian distributions, energy-based score (Liu et al., 2020) aligns better with the probability density. Besides directly designing new score functions, many other works pay attention to various aspects to enhance the OOD detection such that LogitNorm (Wei et al., 2022) produces confidence scores by training with a constant vector norm on the logits, and DICE (Sun & Li, 2022) reduces the variance of the output distribution by leveraging the model sparsification.

OOD Detection with auxiliary data. Another promising direction toward OOD detection involves the auxiliary outliers for model regularization. On the one hand, some works generate virtual outliers such that Lee et al. (2018a) uses generative adversarial networks to generate boundary samples, VOS (Du et al., 2022) regularizes the decision boundary by adaptively sampling virtual outliers from the low-likelihood region. On the other hand, other works tend to exploit information from natural outliers, such that outlier exposure is introduced by Hendrycks et al. (2019b), given that diverse data are available in enormous quantities. (Yu & Aizawa, 2019) train an additional "head" and maximizes the discrepancy of decision boundaries of the two heads to detect OOD samples. Energy-bounded learning (Liu et al., 2020) fine-tunes the neural network to widen the energy gap by adding an energy loss term to the objective. Some other works also highlight the sampling strategy, such that ATOM (Chen et al., 2021) greedily utilizes informative auxiliary data to tighten the decision boundary for OOD detection, and POEM (Ming et al., 2022b) adopts Thompson sampling to contour the decision boundary precisely. The performance of training with outliers is usually superior to that without outliers, shown in many other works (Liu et al., 2020; Fort et al., 2021; Sun et al., 2021; Sehwag et al., 2021; Chen et al., 2021; Salehi et al., 2021; Wei et al., 2022).

6. Conclusion

In this work, we explore the intrinsic OOD discriminative capability of a well-trained model from a unique data-level attribution. Without involving any auxiliary outliers in training, we reveal the inconsistent trend between minimizing original training loss and gaining OOD detection capability. We further identify the potential attribution to be the memorization on atypical samples. To excavate the overlaid capability, we propose the novel Unleashing Mask (UM) and its practical variant UMAP. Through this, we construct model-level discrepancy that figures out the memorized atypical samples and utilizes the constrained gradient ascent to encourage forgetting. It better utilizes the well-trained given model via backtracking or sub-structure pruning. We hope our work could provide new insights for revisiting the model development in OOD detection, and draw more attention toward the data-level attribution. Future work can be extended to a more systematical ID/OOD data investigation with other topics like data pruning or few-shot finetuning.

Acknowledgements

JNZ and BH were supported by NSFC Young Scientists Fund No. 62006202, Guangdong Basic and Applied Basic Research Foundation No. 2022A1515011652, CAAI-Huawei MindSpore Open Fund, and HKBU CSD Departmental Incentive Grant. JCY was supported by the National Key R&D Program of China (No. 2022ZD0160703), STCSM (No. 22511106101, No. 22511105700, No. 21DZ1100100), 111 plan (No. BP0719010). JLX was supported by RGC grants 12202221 and C2004-21GF.

References

Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A. C., Bengio, Y., and Lacoste-Julien, S. A closer look at memorization in deep networks. In *ICML*, 2017.

Barham, P., Chowdhery, A., Dean, J., Ghemawat, S., Hand, S., Hurt, D., Isard, M., Lim, H., Pang, R., Roy, S., et al. Pathways: Asynchronous distributed dataflow for ml. In *MLSys*, 2022.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. In *PNAS*, 2019.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019.

Chen, J., Li, Y., Wu, X., Liang, Y., and Jha, S. Atom: Robustifying out-of-distribution detection using outlier mining. In *ECML PKDD*, 2021.

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

Davis, J. and Goadrich, M. The relationship between precision-recall and roc curves. In *ICML*, 2006.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Djurisic, A., Bozanic, N., Ashok, A., and Liu, R. Extremely simple activation shaping for out-of-distribution detection. In *ICLR*, 2023.

Du, X., Wang, Z., Cai, M., and Li, Y. VOS: learning what you don't know by virtual outlier synthesis. In *ICLR*, 2022.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 2011.

Fang, Z., Li, Y., Lu, J., Dong, J., Han, B., and Liu, F. Is out-of-distribution detection learnable? In *NeurIPS*, 2023.

Fawcett, T. An introduction to roc analysis. In *Pattern Recognition Letters*, 2006.

Fix, E. and Hodges, J. L. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 1989.

Fort, S., Ren, J., and Lakshminarayanan, B. Exploring the limits of out-of-distribution detection. In *NeurIPS*, 2021.

Frankle, J. and Carbin, M. The lottery ticket hypothesis: Training pruned neural networks. In *ICLR*, 2019.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. *Deep learning*. MIT Press, 2016.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Chuang, Z., and Bengio, Y. Challenges in representation learning: A report on three machine learning contests. In *NeurIPS*, 2013.

Han, S., Mao, H., and Dally, W. J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In *ICLR*, 2016.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Tibshirani, R., and Friedman, J. Model assessment and selection. *The elements of statistical learning: data mining, inference, and prediction*, 2009.

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

Hendrycks, D., Lee, K., and Mazeika, M. Using pre-training can improve model robustness and uncertainty. In *ICML*, 2019a.

Hendrycks, D., Mazeika, M., and Dietterich, T. Deep anomaly detection with outlier exposure. In *ICLR*, 2019b.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *CVPR*, 2017.

Huang, R., Geng, A., and Li, Y. On the importance of gradients for detecting distributional shifts in the wild. In *NeurIPS*, 2021.

Ishida, T., Yamane, I., Sakai, T., Niu, G., and Sugiyama, M. Do we need zero training loss after achieving zero training error? In *ICML*, 2020.

Katz-Samuels, J., Nakhleh, J. B., Nowak, R., and Li, Y. Training ood detectors in their natural habitats. In *ICML*, 2022.

Kiefer, J. and Wolfowitz, J. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 1952.

Koltchinskii, V. and Panchenko, D. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 2002.

Krizhevsky, A. Learning multiple layers of features from tiny images. In *arXiv*, 2009.

Lee, K., Lee, H., Lee, K., and Shin, J. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *ICLR*, 2018a.

Lee, K., Lee, K., Lee, H., and Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018b.

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

Lin, Z., Roy, S. D., and Li, Y. Mood: Multi-level out-of-distribution detection. In *CVPR*, 2021.

Liu, W., Wang, X., Owens, J. D., and Li, Y. Energy-based out-of-distribution detection. In *NeurIPS*, 2020.

Loshchilov, I. and Hutter, F. SGDR: stochastic gradient descent with warm restarts. In *ICLR*, 2017.

Manning, C. D. and Schütze, H. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., and Li, Y. Delving into out-of-distribution detection with vision-language representations. In *NeurIPS*, 2022a.

Ming, Y., Fan, Y., and Li, Y. Poem: Out-of-distribution detection with posterior sampling. In *ICML*, 2022b.

Mohseni, S., Pitale, M., Yadawa, J. B. S., and Wang, Z. Self-supervised learning for generalizable out-of-distribution detection. In *AAAI*, 2020.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop*, 2011.

Nguyen, A., Yosinski, J., and Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, 2015.

Ramanujan, V., Wortsman, M., Kembhavi, A., Farhadi, A., and Rastegari, M. What's hidden in a randomly weighted neural network? In *CVPR*, 2020.

Salehi, M., Mirzaei, H., Hendrycks, D., Li, Y., Rohban, M. H., and Sabokrou, M. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. In *arXiv*, 2021.

Sehwag, V., Chiang, M., and Mittal, P. SSD: A unified framework for self-supervised outlier detection. In *ICLR*, 2021.

Sorg, J., Lewis, R. L., and Singh, S. Reward design via online gradient ascent. In *NeurIPS*, 2010.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 2014.

Sun, Y. and Li, Y. Dice: Leveraging sparsification for out-of-distribution detection. In *ECCV*, 2022.

Sun, Y., Guo, C., and Li, Y. React: Out-of-distribution detection with rectified activations. In *NeurIPS*, 2021.

Sun, Y., Ming, Y., Zhu, X., and Li, Y. Out-of-distribution detection with deep nearest neighbors. In *ICML*, 2022.

Tack, J., Mo, S., Jeong, J., and Shin, J. CSI: novelty detection via contrastive learning on distributionally shifted instances. In *NeurIPS*, 2020.

Tavanaei, A. Embedded encoder-decoder in convolutional networks towards explainable AI. In *arXiv*, 2020.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 1933.

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

Vaze, S., Han, K., Vedaldi, A., and Zisserman, A. Open-set recognition: A good closed-set classifier is all you need. In *ICLR*, 2022.

Wang, Q., Ye, J., Liu, F., Dai, Q., Kalander, M., Liu, T., Hao, J., and Han, B. Out-of-distribution detection with implicit outlier transformation. In *ICLR*, 2023.

Wei, H., Xie, R., Cheng, H., Feng, L., An, B., and Li, Y. Mitigating neural network overconfidence with logit normalization. In *ICML*, 2022.

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

Yang, J., Zhou, K., Li, Y., and Liu, Z. Generalized out-of-distribution detection: A survey. In *arXiv*, 2021.

Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., PENG, W., Wang, H., Chen, G., Li, B., Sun, Y., et al. Openood: Benchmarking generalized out-of-distribution detection. In *NeurIPS Datasets and Benchmarks Track*, 2022.

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

Yu, Q. and Aizawa, K. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *ICCV*, 2019.

Zagoruyko, S. and Komodakis, N. Wide residual networks. In *BMVC*, 2016.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.

Zhou, B., Lapedriza, A., Torralba, A., and Oliva, A. Places: An image database for deep scene understanding. *Journal of Vision*, 2017.

Appendix

Reproducibility Statement

We provide the link of our source codes to ensure the reproducibility of our experimental results: `https://github.com/tmlr-group/Unleashing-Mask`. Below we summarize critical aspects to facilitate reproducible results:

- **Datasets.** The datasets we used are all publicly accessible, which is introduced in Section 4.1. For methods involving auxiliary outliers, we strictly follow previous works (Sun et al., 2021; Du et al., 2022) to avoid overlap between the auxiliary dataset (ImageNet-1k) (Deng et al., 2009) and any other OOD datasets.
- **Assumption.** We set our experiments to a post-hoc scenario (Liu et al., 2020) where a well-trained model is available, and some parts of training samples are also available for subsequent fine-tuning (Hendrycks et al., 2019b).
- **Environment.** All experiments are conducted with multiple runs on NVIDIA Tesla V100-SXM2-32GB GPUs with Python 3.6 and PyTorch 1.8.

A. Details about Considered Baselines and Metrics

In this section, we provide the details about the baselines for the scoring functions and fine-tuning with auxiliary outliers, as well as the corresponding hyper-parameters and other related metrics that are considered in our work.

Maximum Softmax Probability (MSP). (Hendrycks & Gimpel, 2017) proposes to use maximum softmax probability to discriminate ID and OOD samples. The score is defined as follows,

$$S_{\text{MSP}}(x; f) = \max_{c} P(y = c|x; f) = \max \texttt{softmax}(f(x)) \tag{6}$$

where f represents the given well-trained model and c is one of the classes $\mathcal{Y} = \{1, \dots, C\}$. The larger softmax score indicates the larger probability for a sample to be ID data, reflecting the model's confidence on the sample.

ODIN. (Liang et al., 2018) designed the ODIN score, leveraging the temperature scaling and tiny perturbations to widen the gap between the distributions of ID and OOD samples. The ODIN score is defined as follows,

$$S_{\text{ODIN}}(x; f) = \max_{c} P(y = c|\tilde{x}; f) = \max \texttt{softmax}(\frac{f(\tilde{x})}{T}) \tag{7}$$

where $\tilde{x}$ represents the perturbed samples (controled by ϵ), T represents the temperature. For fair comparison, we adopt the suggested hyperparameters (Liang et al., 2018): $\epsilon = 1.4 \times 10^{-3}$, $T = 1.0 \times 10^{4}$.

Mahalanobis. (Lee et al., 2018b) introduces a Mahalanobis distance-based confidence score, exploiting the feature space of the neural networks by inspecting the class conditional Gaussian distributions. The Mahalanobis distance score is defined as follows,

$$S_{\text{Mahalanobis}}(x; f) = \max_{c} -(f(x) - \hat{\mu}_c)^T \hat{\Sigma}^{-1} (f(x) - \hat{\mu}_c) \tag{8}$$

where $\hat{\mu}_c$ represents the estimated mean of multivariate Gaussian distribution of class c, $\hat{\Sigma}$ represents the estimated tied covariance of the C class-conditional Gaussian distributions.

Energy. (Liu et al., 2020) proposes to use the Energy of the predicted logits to distinguish the ID and OOD samples. The Energy score is defined as follows,

$$S_{\text{Energy}}(x; f) = -T \log \sum_{c=1}^{C} e^{f(x)_c / T} \tag{9}$$

where T represents the temperature parameter. As theoretically illustrated in Liu et al. (2020), a lower Energy score indicates a higher probability for a sample to be ID. Following (Liu et al., 2020), we fix the T to 1.0 throughout all experiments.

Outlier Exposure (OE). (Hendrycks et al., 2019b) initiates a promising approach towards OOD detections by involving outliers to force apart the distributions of ID and OOD samples. In the experiments, we use the cross-entropy from $f(x_{\text{out}})$to the uniform distribution as the $\mathcal{L}_{\text{OE}}$ (Lee et al., 2018a),

$$\mathcal{L}_f = \mathbb{E}_{\mathcal{D}_{\text{in}}}\left[\ell_{\text{CE}}(f(x), y)\right] + \lambda \mathbb{E}_{\mathcal{D}^{s}_{\text{out}}}\left[\log \sum_{c=1}^{C} e^{f(x)_c} - \mathbb{E}_{\mathcal{D}^{s}_{\text{out}}}(f(x))\right] \tag{10}$$

Energy (w. $\mathcal{D}_{\text{aux}}$). In addition to using the Energy as a post-hoc score to distinguish ID and OOD samples, (Liu et al., 2020) proposes an Energy-bounded objective to further separate the two distributions. The OE objective is as follows,

$$\mathcal{L}_{\text{OE}} = \mathbb{E}_{\mathcal{D}^{s}_{\text{in}}}(\max(0, S_{\text{Energy}}(x, f) - m_{\text{in}}))^2 + \mathbb{E}_{\mathcal{D}^{s}_{\text{out}}}(\max(0, m_{\text{out}} - S_{\text{Energy}}(x, f)))^2 \tag{11}$$

We keep the thresholds same to (Liu et al., 2020): $m_{\text{in}} = -25.0$, $m_{\text{out}} = -7.0$.

POEM. (Ming et al., 2022b) explores the Thompson sampling strategy (Thompson, 1933) to make the most use of outliers to learn a tight decision boundary. Though given the POEM's nature to be orthogonal to other OE methods, we use the Energy(w. $\mathcal{D}_{\text{aux}}$) as the backbone, which is the same as Eq.(11) in Liu et al. (2020). The details of Thompson sampling can refer to Ming et al. (2022b).

FPR and TPR. Suppose we have a binary classification task (to predict an image to be an ID or OOD sample in this paper). There are two possible outputs: a positive result (the model predicts an image to be an ID sample); a negative result (the model predicts an image to be an OOD sample). Since we have two possible labels and two possible outputs, we can form a confusion matrix with all possible outputs as follows,

Table 3. Confusion Matrix.

	Truth: ID	Truth: OOD
Predict: ID	True Positive (TP)	False Positive (FP)
Predict: OOD	False Negative (FN)	True Negative (TN)

The false positive rate (FPR) is calculated as:

$$\text{FPR} = \frac{FP}{FP + TN} \tag{12}$$

The true positive rate (TPR) is calculated as:

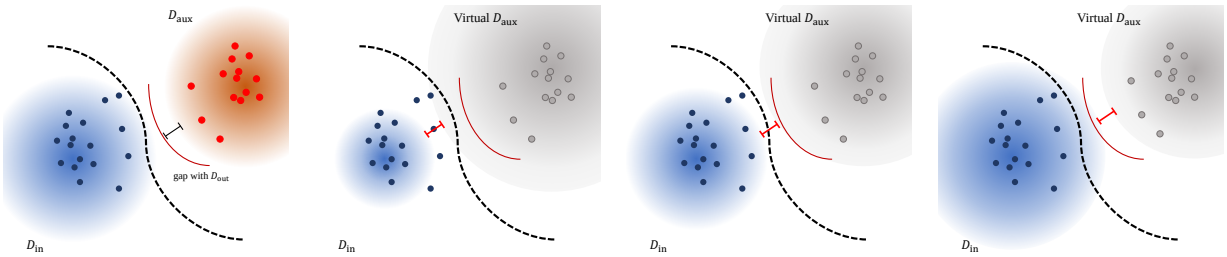
$$\text{TPR} = \frac{TP}{TP + FN} \tag{13}$$

Margin value. Let $f(x) : \mathbb{R}^d \to \mathbb{R}^k$ be a model that outputs k logits, following previous works (Koltchinskii & Panchenko, 2002; Cao et al., 2019), the margin value of an example (x,y) used in our Figure 2(c) is defined as,

$$S_{\text{margin}}(x, y) = f(x)_y - \max_{j \neq y} f(x)_j \tag{14}$$

B. Theoretical Insights on ID Data Property

In this section, we provide a detailed discussion and theoretical analysis to explain the revealed observation and the benefits of our proposed method on ID data property. Specifically, we present the analysis based on the view of sample complexity adopted in POEM (Ming et al., 2022b). To better demonstrate the conceptual extension, we also provide an intuitive illustration based on a comparison with POEM's previous focus on auxiliary outlier sampling in Figure 7 (extended version of Figure 6). Briefly, we focus on the ID data property which is not discussed in the previous analytical framework.



(a) $\mathcal{D}_{\text{in}}$ as an Anchor (b) Under-represent on $\mathcal{D}_{\text{in}}$ (c) Optimal-represent on $\mathcal{D}_{\text{in}}$ (d) Over-represent on $\mathcal{D}_{\text{in}}$

Figure 7. Illustration about the theoretical insights on ID data property considering the binary classification scenario, which is presented as a conceptual comparison based on the underlying intuition in POEM (Ming et al., 2022b). Different from treating the ID distribution $\mathcal{D}_{\text{in}}$ as an analytical anchor in POEM, we present three conceptual visualizations which correspond to different training phases in model development on ID data. (a) Using $\mathcal{D}_{\text{in}}$ as an anchor, the boundary data (defined in Ming et al. (2022b)) sampled from $\mathcal{D}_{\text{aux}}$ is important to mitigate the distribution gap with the true $\mathcal{D}_{\text{out}}$ (as indicated with the red arc). Without the $\mathcal{D}_{\text{aux}}$, we assume a virtual $\mathcal{D}_{\text{aux}}$ exists for the analytical target, which is highly related to the $\mathcal{D}_{\text{in}}$ and the model in the original classification task on $\mathcal{D}_{\text{in}}$. (b) the model under-represent on $\mathcal{D}_{\text{in}}$ with the lower confidence on the atypical samples close to the boundary; (c) the model reaches a near-optimal representation status on $\mathcal{D}_{\text{in}}$; (d) the model over-represent on $\mathcal{D}_{\text{in}}$. The corresponding OOD discriminative capability is affected by the different scenarios.

Preliminary setup and notations. As the original training task (e.g., the multi-classification task on CIFAR-10) does not involve any outliers data, it is hard to analyze the related property with OOD detection. Here we introduce an Assumption B.1 about virtual $\mathcal{D}_{\text{aux}}$ to help complete the analytical framework. To sum up, we consider a binary classification task here for distinguishing ID and OOD data. Following the prior works (Lee et al., 2018a; Sehwag et al., 2021; Ming et al., 2022b), we assume the extracted feature approximately follows a Gaussian mixture model (GMM) with the equal class priors as $\frac{1}{2}\mathcal{N}(\mu, \sigma^2\mathcal{I}) + \frac{1}{2}\mathcal{N}(-\mu, \sigma^2\mathcal{I})$. To be specific, $\mathcal{D}_{\text{in}} = \mathcal{N}(\mu, \sigma^2\mathcal{I})$ and $\mathcal{D}_{\text{aux}} = \mathcal{N}(-\mu, \sigma^2\mathcal{I})$. Considering the hypothesis class as $\mathcal{H} = \text{sign}(\theta^T x), \theta \in \mathbb{R}^d$. The classifier outputs 1 if $x \sim \mathcal{D}_{\text{in}}$ and outputs -1 if $x \sim \mathcal{D}_{\text{aux}}$.

First, we introduce the assumption about virtual $\mathcal{D}_{\text{aux}}$. Considering the representation power of deep neural networks, the assumption can be valid. It is empirically supported by the evidence in Figure 6, as the part of real $\mathcal{D}_{\text{out}}$ can be viewed as the virtual $\mathcal{D}_{\text{aux}}$. Second, to better link our method for the analysis, we introduce another assumption (i.e., Assumption B.2) about the ID training status. It can be verified by the relative degree of distinguishability indicated by a fixed threshold in Figure 6, that the model is more confident on the $\mathcal{D}_{\text{out}}$ along with the training.

Assumption B.1 (Virtual $\mathcal{D}_{\text{aux}}$)**.** Given the well-trained model in the original classification task on the ID distribution $\mathcal{D}_{\text{in}}$, and considering the binary classification for OOD detection, we can assume the existence of a virtual $\mathcal{D}_{\text{aux}}$, that the OOD discriminative capacity of the current model can result from learning on the virtual $\mathcal{D}_{\text{aux}}$ with the outlier exposure manner.

Assumption B.2 (ID Training Status w.r.t. Masking)**.** Considering the model training phase in the original multi-class classification task on the ID distribution $\mathcal{D}_{\text{in}}$, and tuning with a specific mask ratio serving as the sample selection, we assume that the data points $x \sim$ virtual $\mathcal{D}_{\text{aux}}$ satisfy the extended constraint based on the boundary scores $-|f_{\text{outlier}}(x)|$ defined in POEM (Ming et al., 2022b): $\sum_{i=1}^{n} f_{\text{outlier}} \leq (|\alpha - \delta^*| + \epsilon)n$, where the f_{outlier} is a function parameterized by some unknown ground truth weights and maps the high-dimensional input x into a scalar. Generally, it represents the discrepancy between virtual $\mathcal{D}_{\text{aux}}$ and the true $\mathcal{D}_{\text{out}}$, indicated with the constraint $|\alpha - \delta^*|$ results from the masked ID data.

Given the above, we can naturally get the following extended lemma based on that adopted in POEM (Ming et al., 2022b).

Lemma B.3 (Constraint of Varied Virtual $\mathcal{D}_{\text{aux}}$)**.** *Assume the data points $x\sim$ virtual $\mathcal{D}_{aux}$ satisfy the following constraint for resulting in the following varied boundary margin: $\sum_{i=1}^{n} |2x_i^T \mu| \leq n\sigma^2(|\alpha - \delta^*| + \epsilon)$.*

proof of Lemma. B.3. Given the Gaussian mixture model described in the previous setup, we can obtain the following expression by Bayes' rule of $\mathbb{P}(\text{outlier}|x)$,

$$\mathbb{P}(\text{outlier}|x) = \frac{\mathbb{P}(x|\text{outlier})\mathbb{P}(\text{outlier})}{\mathbb{P}(x)} = \frac{1}{1 + e^{-\frac{1}{2\sigma^2}(d_{\text{outlier}}(x) - d_{\text{in}}(x))}}, \tag{15}$$

where $d_{\text{outlier}}(x)) = (x+\mu)^\top(x+\mu)$, $d_{\text{in}}(x) = (x-\mu)^\top(x-\mu)$, and $\mathbb{P}(\text{outlier}|x) = \frac{1}{1+e^{-f_{\text{outlier}}(x)}}$ according to its definition.

Then we have:

$$-f_{\text{outlier}} = -\frac{1}{2\sigma^2}(d_{\text{outlier}}(x) - d_{\text{in}}(x)), \tag{16}$$

$$-|f_{\text{outlier}}| = -\frac{1}{2\sigma^2}|(x-\mu)^\top(x-\mu) - (x+\mu)^\top(x+\mu)| = -\frac{2}{\sigma^2}|x^\top\mu|. \tag{17}$$

Therefore, we can get the constraint as: $\sum_{i=1}^{n}|2x_i^T\mu| \leq n\sigma^2(|\alpha-\delta^*|+\epsilon)$. □

With the previous assumption and lemma that incorporate our masking in the variable δ^*, we present the analysis as below.

Complexity analysis anchored on $\mathcal{D}_{\text{in}}$. With the above lemma and the assumptions of virtual $\mathcal{D}_{\text{aux}}$ (as illustrated in Figure 7), we can derive the results to understand the benefits from the revealed observation and our UM and UMAP.

Consider the given classifier defined as $\theta^*_{n_1,n_2} = \frac{1}{n_1+n_2}(\sum_{i=1}^{n_1}x_i^1 - \sum_{i=1}^{n_2}x_i^2)$, assume each x_i^1 is drawn *i.i.d.* from $\mathcal{D}_{\text{in}}$ and each x_i^2 is drawn *i.i.d* from $\mathcal{D}_{\text{aux}}$, and assume the signal/noise ratio is $\frac{||\mu||}{\sigma} = r_0 \gg 1$, the dimentionality/sample size ratio is $\frac{d}{n} = r_1$, as well as exist some constant $\alpha < 1$. By decomposition, we can rewrite $\theta^*_{n_1,n_2} = \mu + \frac{n_1}{n_1+n_2}\theta_1 + \frac{n_2}{n_1+n_2}\theta_2$ with the following θ_1 and θ_2:

$$\theta_1 = \frac{1}{n_1}(\sum_{i=1}^{n_1}x_i^1) - \mu, \quad \theta_2 = \frac{1}{n_2}(-\sum_{i=1}^{n_2}x_i^2) - \mu, \tag{18}$$

Since $\theta_1 \sim \mathcal{N}(0, \frac{\sigma^2}{n_1}\mathcal{I})$, we have that $||\theta_1||^2 \sim \frac{\sigma^2}{n_1}\mathcal{X}_d^2$ and $\frac{\mu^T\theta_1}{||\mu||} \sim \mathcal{N}(0, \frac{\sigma^2}{n_1})$ to form the standard concentration bounds as:

$$\mathbb{P}(||\theta_1||^2 \geq \frac{\sigma^2}{n_1}(d + \frac{1}{\sigma})) \leq e^{-\frac{d}{8\sigma^2}}, \quad \mathbb{P}(\frac{|\mu^T\theta_1|}{||\mu||} \geq (\sigma||\mu||)^{\frac{1}{2}}) \leq 2e^{-\frac{n_1||\mu||}{2\sigma}} \tag{19}$$

Anchored on $\mathcal{D}_{\text{in}}$, the distribution of θ_2 can be treated as a truncated distribution of θ_1 as x_i^2 drawn *i.i.d.* from the virtual $\mathcal{D}_{\text{aux}}$ are under the relative constraint with $\mathcal{D}_{\text{in}}$. Without losing the generality, we replace n_1 with n, and have the following inequality with a finite positive constant a:

$$\mathbb{P}(||\theta_2||^2 \geq \frac{\sigma^2}{n_1}(d + \frac{1}{\sigma})) \leq ae^{-\frac{d}{8\sigma^2}} \tag{20}$$

According to Lemma B.3, we can have that $|\mu^T\theta_2| \leq ||\mu||^2 + \frac{\sigma^2(|\alpha-\delta^*|+\epsilon)}{2}$. Now we can have $||\theta_1||^2 \leq \frac{\sigma^2}{n}(d+\frac{1}{\sigma})$, $||\theta_2||^2 \leq \frac{\sigma^2}{n}(d+\frac{1}{\sigma})$, $\frac{|\mu^T\theta_1|}{||\mu||} \leq (\sigma||\mu||)^{\frac{1}{2}}$ simultaneously hold and derive the following recall the decomposition,

$$||\theta^*_{n_1,n_2}||^2 = ||\mu + \frac{n_1}{n_1+n_2}\theta_1 + \frac{n_2}{n_1+n_2}\theta_2||^2 \leq \frac{\sigma^2}{n}(d+\frac{1}{\sigma}) + ||\mu||^2, \tag{21}$$

and

$$|\mu^T\theta^*_{n_1,n_2}| \geq \frac{1}{2}(||\mu||^2 - \sigma^{\frac{1}{2}}||\mu||^{\frac{3}{2}} - \frac{\sigma^2(|\alpha-\delta^*|+\epsilon)}{2}). \tag{22}$$

With the above inequality derived in Eq. (21) and Eq. (22), we can have the following bound with the probability at least $1-(1+a)e^{-\frac{r_1 n}{8\sigma^2}} - 2e^{-\frac{n_1||\mu||}{2\sigma}}$

$$\frac{\mu^T\theta^*_{n_1,n_2}}{\sigma||\theta^*_{n_1,n_2}||} \geq \frac{||\mu||^2 - \sigma^{\frac{1}{2}}||\mu||^{\frac{3}{2}} - \frac{\sigma^2(|\alpha-\delta^*|+\epsilon)}{2}}{2\sqrt{\frac{\sigma^2}{n}(d+\frac{1}{\sigma}) + ||\mu||^2}} \tag{23}$$

Since $\text{FPR}(\theta^*_{n_1,n_2}) = erf(\frac{\mu^T\theta^*_{n_1,n_2}}{\sigma||\theta^*_{n_1,n_2}||})$ is monotonically decreasing, as the lower bound of $\frac{\mu^T\theta^*_{n_1,n_2}}{\sigma||\theta^*_{n_1,n_2}||}$ will increase as the constraint from the virtual $\mathcal{D}_{\text{aux}}$ changed accordingly in our UM and UMAP, the upper bound of $\text{FPR}(\theta^*_{n_1,n_2})$ will decrease. From the above analysis, one insight we can draw is learning more atypical ID data may need more high-quality auxiliary outliers to shape the near-the-boundary behavior of the model, which can further enhance the OOD discriminative capability.

C. Discussion about the "Conflict" Against Previous Empirical Observation

In this section, we address what initially appears to be a contradiction between our observation and previous empirical studies (Vaze et al., 2022; Fort et al., 2021), but it is not a contradiction. This work demonstrates that during training, there exists a middle stage where the model's OOD detection performance is superior to the final stage, even though the model has not achieved the best performance on ID-ACC. Some previous studies (Vaze et al., 2022; Fort et al., 2021) suggest that a good close-set classifier tends to have higher OOD detection performance, which may seem to contradict our claim. However, this is not the case, and we provide the following explanations.

First, the previous empirical observation (Vaze et al., 2022; Fort et al., 2021) of a high correlation between a good close-set classifier (e.g., high ID-ACC in (Vaze et al., 2022)) and OOD detection performance is based on **inter-model comparisons**, such as comparing different model architectures. This is consistent with our results in Table 4. Even the previous model stages backtracked via our UM show similar results confirming that a better classifier (e.g., the DenseNet-101 in Table 4) is better to achieve better OOD detection performance.

Second, our observation is based on **intra-model comparisons**, which compare different training stages of a single model. Our results in Figure 1 across various training settings confirm this observation. Additionally, Table 5 shows that when we backtrack the model through UM, we obtain lower ID-ACC but better OOD detection performance. However, if we compare different models, DenseNet-101 with higher ID-ACC still outperforms Wide-ResNet, as previously mentioned.

To summarize, our observation provides an orthogonal view to exploring the relationship between ID-ACC and OOD detection performance. On the one hand, we attribute this observation to the model's memorization of atypical samples, as further demonstrated by our experiments (e.g., in Figure 2). On the other hand, we believe that this observation reveals other characteristics of a "good classifier" beyond ID-ACC, e.g., higher OOD detection capability.

Table 4. **Inter-model comparison** (different models) of ID-ACC with the OOD detection performance on CIFAR-10 (%). ↑ indicates higher values are better, and ↓ indicates lower values are better.

Method	Model	AUROC↑	AUPR↑	FPR95↓	ID-ACC↑
MSP	DenseNet-101	**89.90 ± 0.30**	**91.48 ± 0.43**	**60.08 ± 0.76**	**94.01 ± 0.08**
	WRN-40-4	87.12 ± 0.25	87.84 ± 0.30	68.29 ± 0.96	93.86 ± 0.19
ODIN	DenseNet-101	**91.46 ± 0.56**	**91.67 ± 0.58**	**42.31 ± 1.38**	**94.01 ± 0.08**
	WRN-40-4	83.29 ± 0.72	82.74 ± 0.79	65.68 ± 0.77	93.86 ± 0.19
Energy	DenseNet-101	**92.07 ± 0.22**	**92.72 ± 0.39**	**42.69 ± 1.31**	**94.01 ± 0.08**
	WRN-40-4	87.69 ± 0.54	88.16 ± 0.69	58.47 ± 1.94	93.86 ± 0.19
Energy+**UM** (ours)	DenseNet-101	**93.73 ± 0.36**	**94.27 ± 0.60**	**33.29 ± 1.70**	**92.80 ± 0.47**
	WRN-40-4	91.74 ± 0.43	92.67 ± 0.52	40.40 ± 1.32	92.68 ± 0.23

Table 5. **Intra-model comparison** (regarding the same model) of ID-ACC with the OOD detection performance on CIFAR-10 (%). ↑ indicates higher values are better, and ↓ indicates lower values are better.

Method	Model	AUROC↑	AUPR↑	FPR95↓	ID-ACC↑
Energy	DenseNet-101	92.07 ± 0.22	92.72 ± 0.39	42.69 ± 1.31	**94.01 ± 0.08**
Energy+**UM** (ours)		**93.73 ± 0.36**	**94.27 ± 0.60**	**33.29 ± 1.70**	92.80 ± 0.47
Energy	WRN-40-4	87.69 ± 0.54	88.16 ± 0.69	58.47 ± 1.94	**93.86 ± 0.19**
Energy+**UM** (ours)		**91.74 ± 0.43**	**92.67 ± 0.52**	**40.40 ± 1.32**	92.68 ± 0.23

D. Discussion with Conventional Overfitting

In this section, we provide a comprehensive comparison of our observation and conventional overfitting in deep learning.

First, we would refer to the concept of the conventional overfitting (Goodfellow et al., 2016; Belkin et al., 2019), i.e., the model "overfits" the training data but fails to generalize and perform well on the test data that is unseen during training.

The common empirical reflection of overfitting is that the training error is decreasing while the test error is increasing at the same time, which enlarges the generalization gap of the model. It has been empirically confirmed not the case in our observation as observed in Figure 2(a) and 2(b). To be specific, for the original classification task, there is no conventional overfitting observed as the test performance is still improved at the later training stage, which is a general pursuit of the model development phase on the original tasks (Goodfellow et al., 2016; Zhang et al., 2017).

Then, when we consider the OOD detection performance of the well-trained model, our unique observation is about the inconsistency between gaining better OOD detection capability and pursuing better performance on the original classification task for the in-distribution (ID) data. It is worth noting that here the training task is not the binary classification of OOD detection, but the classification task on ID data. It is out of the rigorous concept of conventional overfitting and has received limited focus and discussion through the data-level perspective in the previous literature about OOD detection (Yang et al., 2021; 2022) to the best of our knowledge. Considering the practical scenario that exists target-level discrepancy, our revealed observation may encourage us to revisit the detection capability of the well-trained model.

Third, we also provide an empirical comparison with some strategies targeted for mitigating overfitting. In our experiments, for all the baseline models including that used in Figure 1, we have adopted those strategies (Srivastava et al., 2014; Hastie et al., 2009) (e.g., drop-out, weight decay) to reduce overfitting. The results are summarized in the following Tables 6, 7, 8 and 9. According to the experiments, most conventional methods proposed to prevent conventional overfitting show limited benefits in gaining better OOD detection performance, since they have a different underlying target from UM/UMAP. However, most of them suffer from the higher sacrifice on the performance of the original task and may not be compatible and practical in the current general setting, i.e., starting from a well-trained model. In contrast, our proposed UMAP can be a more practical and flexible way to restore detection performance.

Table 6. Comparison among overfitting methods and ODIN with DenseNet-101 (%). ↑ indicates higher values are better, and ↓ indicates lower values are better.

$\mathcal{D}_{\text{in}}$	Method	AUROC↑	AUPR↑	FPR95↓	ID-ACC↑
CIFAR-10	Baseline	91.67	91.89	40.74	93.67
	Early Stopping w. ACC	92.13	92.46	38.86	93.69
	Weight Decay 0.1	86.64	86.67	60.07	88.53
	Weight Decay 0.01	90.76	91.25	44.20	92.07
	Weight Decay 0.001	88.93	88.25	48.95	94.26
	Drop Rate 0.3	91.14	92.21	46.58	90.05
	Drop Rate 0.4	84.95	86.62	62.52	82.55
	Drop Rate 0.5	83.75	85.17	62.17	75.31
	UM (ours)	92.45	93.06	37.13	92.76
	UMAP (ours)	91.92	92.88	37.69	93.69

Table 7. Comparison among overfitting methods and Energy with DenseNet-101 (%). ↑ indicates higher values are better, and ↓ indicates lower values are better.

$\mathcal{D}_{\text{in}}$	Method	AUROC↑	AUPR↑	FPR95↓	ID-ACC↑
CIFAR-10	Baseline	92.72	93.48	38.30	93.67
	Early Stopping w. ACC	92.75	93.54	37.84	93.69
	Weight Decay 0.1	86.78	88.04	65.08	88.53
	Weight Decay 0.01	90.86	91.77	47.64	92.07
	Weight Decay 0.001	90.68	90.90	47.38	94.26
	Drop Rate 0.3	90.52	91.79	51.23	90.05
	Drop Rate 0.4	84.29	86.43	68.17	82.55
	Drop Rate 0.5	83.29	85.14	68.17	75.31
	UM (ours)	93.58	94.14	33.66	92.76
	UMAP (ours)	93.17	93.87	36.11	93.69

Given the concept discrepancy aforementioned, we can know that "memorization of the atypical samples" are not "memorization in overfitting". Those atypical samples are empirically beneficial in improving the performance on the original classification task as shown in Figure 2. However, this part of knowledge is not very necessary and even harmful to the OOD

Table 8. Comparison among overfitting methods and ODIN with WRN-40-4 (%). ↑ indicates higher values are better, and ↓ indicates lower values are better.

$\mathcal{D}_{\text{in}}$	Method	AUROC↑	AUPR↑	FPR95↓	ID-ACC↑
CIFAR-10	Baseline	86.24	85.90	60.13	93.86
	Early Stopping w. ACC	83.80	83.30	65.13	93.99
	Weight Decay 0.1	84.38	84.75	65.75	89.88
	Weight Decay 0.01	88.08	88.45	55.16	93.16
	Weight Decay 0.001	86.34	86.38	57.42	94.91
	Drop Rate 0.3	87.53	87.25	56.12	94.22
	Drop Rate 0.4	88.24	88.41	54.62	94.20
	Drop Rate 0.5	89.13	89.99	53.07	93.91
	UM (ours)	89.61	91.13	50.97	92.68
	UMAP (ours)	90.43	91.73	46.96	93.86

Table 9. Comparison among overfitting methods and Energy with WRN-40-4 (%). ↑ indicates higher values are better, and ↓ indicates lower values are better.

$\mathcal{D}_{\text{in}}$	Method	AUROC↑	AUPR↑	FPR95↓	ID-ACC↑
CIFAR-10	Baseline	87.69	88.16	58.47	93.86
	Early Stopping w. ACC	88.07	88.65	67.61	93.99
	Weight Decay 0.1	86.97	88.51	63.54	89.88
	Weight Decay 0.01	89.77	89.82	50.23	93.16
	Weight Decay 0.001	89.25	89.84	50.95	93.91
	Drop Rate 0.3	89.74	90.07	52.16	93.22
	Drop Rate 0.4	89.94	90.53	51.13	94.20
	Drop Rate 0.5	90.09	91.04	52.76	93.91
	UM (ours)	91.74	92.67	40.40	92.68
	UMAP (ours)	88.84	89.31	50.23	93.86

detection task as the detection performance of the model drops significantly. Based on the training and test curves in our observation, the memorization in overfitting is expected to happen later than the final stage in which the test performance would drop. Since we have already used some strategies to prevent overfitting, it does not exist. Intuitively, the "atypical samples" identified in our work are relative to the OOD detection task. The memorization of "atypical samples" indicates that the model may not be able to draw the general information of the ID distribution through further learning on those atypical samples through the original classification task. Since we mainly provide the understanding of the data-level attribution for OOD discriminative capability, further analysis from theoretical views (Fang et al., 2023) to link the conventional overfitting with OOD detection would be an interesting future direction.

E. Additional Explanation Towards Mining the Atypical Samples

In this section, we provide further discussion and explanation about mining the atypical samples.

First, for identifying those atypical samples using a randomly initialized layer-wise mask (Ramanujan et al., 2020) with the well-pre-trained model, the underlying intuition is constructing the parameter-level discrepancy to mine the atypical samples. It is inspired by and based on the evidence drawn from previous literature about learning behaviors (Arpit et al., 2017; Goodfellow et al., 2016) of deep neural networks (DNNs), sparse representation (Frankle & Carbin, 2019; Goodfellow et al., 2013; Barham et al., 2022), and also model uncertainty representation (like dropout (Gal & Ghahramani, 2016)). To be specific, the atypical samples tend to be learned by the DNNs later than those typical samples (Arpit et al., 2017), and are relatively more sensitive to the changes of the model parameter as the model does not generalize well on that (Srivastava et al., 2014; Gal & Ghahramani, 2016). By the layer-wise mask, the constructed discrepancy can make the model misclassify the atypical samples and estimate loss constraint for the forgetting objective, as visualized in Figure 3.

Second, introducing the layer-wise mask has several advantages for achieving the staged target of mining atypical samples in our proposed method, while we would also admit that the layer-wise mask may not be an irreplaceable option or may not be

optimal. On the one hand, considering that the model has been trained to approach the zero error on training data, utilizing the layer-wise mask is an integrated strategy to 1) figure out the atypical samples; and 2) obtain the loss value computed by the masked output that misclassifies them. The loss constraint is later used in the forgetting objective to fine-tune the model. On the other hand, the layer-wise mask is also compatible with the proposed UMAP to generate a flexible mask for restoring the detection capability of the original model.

More discussion and visualization using CIFAR-10 and ImageNet. Third, we also adopt the unit/weight mask (Han et al., 2016) and visualize the misclassified samples in Figure 18 (we also present a similar visualization about the experiments on ImageNet (Deng et al., 2009) in Figure 19). The detected samples show that traditionally pruning the network according to weights can't efficiently figure out whether an image is typical or atypical while pruning randomly can do so. Intuitively, we attribute this phenomenon to the uncertain relationship (Gal & Ghahramani, 2016) between the magnitudes and the learned patterns. Randomly masking out weights can have a harsh influence on atypical samples, which creates a discrepancy in mining them. Further investigating the specific effect of different methods that construct the parameter-level discrepancy would be an interesting sub-topic in future work. For the value of CE loss, although the atypical samples tend to have high CE loss value, they are already memorized and correctly classified as indicated by the zero training error. Only using the high CE error can not provide the loss estimation when the model does not correctly classify those samples.

F. Algorithmic Realization of UM and UMAP

In this section, we provide the detailed algorithmic realizations of our proposed Unleashing Mask (UM) (i.e., in Algorithm 1) and Unleashing Mask Adopt Pruning (UMAP) (i.e., in Algorithm 2) given the well-trained model.

In general, we seek to unleash the intrinsic detection power of the well-trained model by adjusting the well-trained given model. For the first part, we need to mine the atypical samples and estimate the loss value to misclassify them using the current model. For the second part, we need to tune or prune with the loss constraint for forgetting.

To estimate the loss constrain for forgetting (i.e., $\widehat{\ell}_{\text{CE}}(m_\delta \odot f^*)$ in Eq 3 with the fixed given model f^*), we randomly knock out parts of weights according to a specific mask ratio δ. To be specific, we sample a score from a Gaussian distribution for every weight. Then we initialize a unit matrix for every layer of the model concerning the size of the layer. We formulate the mask m_δ according to the sampled scores. We then iterate through every layer (termed as $l \in \theta_{\text{layers}}$) to find the threshold for each layer that is smaller than the score of the given mask ratio in that layer (termed as quantile). Then set all the ones, whose corresponding scores are more significant than the layers' thresholds, to zeros.

We dot-multiply every layer's weights with the formulated binary matrix as if we delete some parts of the weights. Then, we input a batch of training samples to the masked model and treat the mean value of the outputs' CE loss as the loss constraint. After all of these have been done, we begin to fine-tune the model's weights with the loss constraint applied to the original CE loss. In our algorithms, the fine-tuning epochs k is the epochs we finetune after we get the well-trained model.

For UMAP, the major difference from UM is that, instead of fine-tuning the weights, we generate a popup score for every weight, and force the gradients to pass through the scores. In every iteration, we need to formulate a binary mask according to the given prune rate p. This is just what we do when estimating the loss constraint. For more details, it can refer to (Ramanujan et al., 2020). In Table 10, we summarize the complete comparison of UM and UMAP to show their effectiveness. We also provide the performance comparison by switching ID training data to be the large-scaled ImageNet, and demonstrate the effectiveness of our UM and UMAP in Table 11. In practice, we use SVHN as a validation OOD set to tune the mask ratio. We adopt 99.6% (the corresponding estimated loss constraint is about 0.6) in this large-scale experiment to estimate the loss constraint for forgetting. Surprisingly, we also find that loss values smaller than the estimated one (i.e., < 0.6) can also help improve OOD detection performance, distinguishing the general effectiveness of UM/UMAP.

Algorithm 1 Unleashing Mask (UM)

Input: well-trained model : θ, Gaussian distribution: $N(\mu, \sigma^2)$, mask ratio : $\delta \in [0, 1]$, fine-tuning epochs of UM : k, training samples : $x \sim \mathcal{D}_{\text{in}}^{\text{s}}$, layer-iterated model : θ_{layer}, compute the δ-th quantile of the data s : quantile(s, δ);
Output: fine-tuned model θ^k;

1: `// Initialize a popup score for every weight`
2: **for** $w \in \theta$ **do**
3: $\quad s_w \sim N(\mu, \sigma^2)$
4: **end for**
5: `// Generate mask by the popup scores`
6: **for** $l \in \theta_{\text{layers}}$ **do**
7: $\quad m_\delta^l = s_l >$ quantile(s_l, δ) `// Generated mask for layer` l
8: **end for**
9: `// Unleashing Mask: fine-tuning`
10: **for** $t \in (1, \ldots, k)$ **do**
11: $\quad \theta^{(t+1)} = \theta^{(t)} - \eta \frac{\partial(|\mathcal{L}_{\text{CE}}(x,\theta^{(t)}) - \mathbb{E}_{x\sim\mathcal{D}_{\text{in}}^{\text{s}}}(\hat{\mathcal{L}}_{\text{CE}}(x, m_\delta \odot \theta^{(t)}))| + \mathbb{E}_{x\sim\mathcal{D}_{\text{in}}^{\text{s}}}(\hat{\mathcal{L}}_{\text{CE}}(x, m_\delta \odot \theta^{(t)})))}{\partial \theta^{(t)}}$
12: **end for**

Algorithm 2 Unleashing Mask Adopt Pruning (UMAP)

Input: well-trained model : θ, Gaussian distribution: $N(\mu, \sigma^2)$, mask ratio: $\delta \in [0, 1]$, fine-tuning epochs of UM: k, training samples: $x \sim \mathcal{D}_{\text{in}}^{\text{s}}$, prune rate: p , layer-iterated model : θ_{layer},compute the δ-th quantile of the data s : quantile(s, δ);
Output: learnt binary mask $\hat{m}_p$;

1: `// Initialize a popup score for every weight`
2: **for** $w \in \theta$ **do**
3: $\quad s_w \sim N(\mu, \sigma^2)$
4: **end for**
5: `// Generate mask by the popup scores`
6: **for** $l \in \theta_{\text{layers}}$ **do**
7: $\quad m_\delta^l = s^l >$ quantile(s^l, δ) `// Generated mask for layer` l
8: **end for**
9: `// Initialize the ready-to-be-learned scores for UMAP`
10: **for** $w \in \theta$ **do**
11: $\quad \hat{s}_w^{(1)} \sim N(\mu, \sigma^2)$
12: **end for**
13: `// Unleashing Mask Adopt Pruning: Pruning`
14: **for** $t \in (1, \ldots, k)$ **do**
15: $\quad$ `// Generate the mask for UMAP according to learned scores` $\hat{s}^t$
16: $\quad$ **for** $l \in \theta_{\text{layers}}$ **do**
17: $\quad\quad \hat{m}_p^l = \hat{s}_l^t >$ quantile$(\hat{s}_l^t, p)$
18: $\quad$ **end for**
19: $\quad \hat{s}^{(t+1)} = \hat{s}^{(t)} - \eta \frac{\partial(|\mathcal{L}_{\text{CE}}(x, \hat{m}_p \odot \theta) - \mathbb{E}_{x \sim \mathcal{D}_{\text{in}}^{\text{s}}}(\hat{\mathcal{L}}_{\text{CE}}(x, m_\delta \odot \theta))| + \mathbb{E}_{x \sim \mathcal{D}_{\text{in}}^{\text{s}}}(\hat{\mathcal{L}}_{\text{CE}}(x, m_\delta \odot \theta)))}{\partial \hat{s}^{(t)}}$
20: **end for**
21: **for** $l \in \theta_{\text{layers}}$ **do**
22: $\quad \hat{m}_p^l = \hat{s}_l^k >$ quantile$(\hat{s}_l^k, p)$
23: **end for**

Table 10. Completed Results (%). Comparison with competitive OOD detection baselines. ↑ indicates higher values are better, and ↓ indicates lower values are better.

$\mathcal{D}_{\text{in}}$	Method	AUROC↑	AUPR↑	FPR95↓	ID-ACC↑	w./w.o $\mathcal{D}_{\text{aux}}$
CIFAR-10	MSP(Hendrycks & Gimpel, 2017)	89.90 ± 0.30	91.48 ± 0.43	60.08 ± 0.76	$\mathbf{94.01 \pm 0.08}$	
	ODIN(Liang et al., 2018)	91.46 ± 0.56	91.67 ± 0.58	42.31 ± 1.38	$\mathbf{94.01 \pm 0.08}$	
	Mahalanobis(Lee et al., 2018b)	75.10 ± 1.04	72.32 ± 1.92	61.35 ± 1.25	$\mathbf{94.01 \pm 0.08}$	
	Energy(Liu et al., 2020)	92.07 ± 0.22	92.72 ± 0.39	42.69 ± 1.31	$\mathbf{94.01 \pm 0.08}$	
	Energy+UM (ours)	93.73 ± 0.36	94.27 ± 0.60	33.29 ± 1.70	92.80 ± 0.47	
	Energy+UMAP (ours)	$\mathbf{93.97 \pm 0.11}$	$\mathbf{94.38 \pm 0.06}$	$\mathbf{30.71 \pm 1.94}$	$\mathbf{94.01 \pm 0.08}$	
	OE(Hendrycks et al., 2019b)	97.07 ± 0.01	97.31 ± 0.05	13.80 ± 0.28	92.59 ± 0.32	✓
	Energy (w. $\mathcal{D}_{\text{aux}}$)(Liu et al., 2020)	94.58 ± 0.64	94.69 ± 0.65	18.79 ± 2.31	80.91 ± 3.13	✓
	POEM(Ming et al., 2022b)	94.37 ± 0.07	94.51 ± 0.06	18.50 ± 0.33	77.24 ± 2.22	✓
	OE+UM (ours)	$\mathbf{97.60 \pm 0.03}$	$\mathbf{97.87 \pm 0.02}$	$\mathbf{11.22 \pm 0.16}$	$\mathbf{93.66 \pm 0.12}$	✓
	Energy+UM (ours)	93.02 ± 0.42	92.36 ± 0.38	24.41 ± 1.65	71.97 ± 0.92	✓
	POEM+UM (ours)	93.04 ± 0.02	92.99 ± 0.02	23.52 ± 0.16	67.41 ± 0.27	✓
	OE+UMAP (ours)	97.48 ± 0.01	97.74 ± 0.00	12.21 ± 0.09	93.44 ± 0.21	✓
	Energy+UMAP (ours)	95.63 ± 1.15	95.92 ± 1.17	17.51 ± 2.59	88.12 ± 4.22	✓
	POEM+UMAP (ours)	94.18 ± 2.98	94.15 ± 3.46	20.55 ± 8.70	76.62 ± 17.95	✓
CIFAR-100	MSP(Hendrycks & Gimpel, 2017)	74.06 ± 0.69	75.37 ± 0.73	83.14 ± 0.87	$\mathbf{74.86 \pm 0.21}$	
	ODIN(Liang et al., 2018)	76.18 ± 0.14	76.49 ± 0.20	78.93 ± 0.31	$\mathbf{74.86 \pm 0.21}$	
	Mahalanobis(Lee et al., 2018b)	63.90 ± 1.91	64.31 ± 0.91	78.79 ± 0.50	$\mathbf{74.86 \pm 0.21}$	
	Energy(Liu et al., 2020)	$\mathbf{76.29 \pm 0.24}$	$\mathbf{77.06 \pm 0.55}$	78.46 ± 0.06	$\mathbf{74.86 \pm 0.21}$	
	Energy+UM (ours)	76.22 ± 0.42	76.39 ± 1.03	74.05 ± 0.55	64.55 ± 0.24	
	Energy+UMAP (ours)	75.57 ± 0.59	75.66 ± 0.07	$\mathbf{72.21 \pm 1.46}$	$\mathbf{74.86 \pm 0.21}$	
	OE(Hendrycks et al., 2019b)	90.55 ± 0.87	90.34 ± 0.94	34.73 ± 3.85	73.59 ± 0.30	✓
	Energy (w. $\mathcal{D}_{\text{aux}}$)(Liu et al., 2020)	88.92 ± 0.57	89.13 ± 0.56	37.90 ± 2.59	57.85 ± 2.65	✓
	POEM(Ming et al., 2022b)	88.95 ± 0.54	88.94 ± 0.31	38.10 ± 1.30	56.18 ± 1.92	✓
	OE+UM (ours)	91.04 ± 0.11	91.13 ± 0.24	34.71 ± 0.81	$\mathbf{75.15 \pm 0.18}$	✓
	Energy+UM (ours)	90.39 ± 0.40	90.14 ± 0.45	32.65 ± 3.13	71.95 ± 0.23	✓
	POEM+UM (ours)	$\mathbf{91.18 \pm 0.35}$	$\mathbf{91.45 \pm 0.27}$	$\mathbf{30.78 \pm 1.76}$	70.17 ± 0.01	✓
	OE+UMAP (ours)	91.10 ± 0.16	90.99 ± 0.23	33.62 ± 0.26	74.76 ± 0.11	✓
	Energy+UMAP (ours)	90.52 ± 0.26	90.46 ± 0.50	32.17 ± 0.30	72.76 ± 0.18	✓
	POEM+UMAP (ours)	91.10 ± 0.29	91.41 ± 0.28	31.02 ± 1.70	71.05 ± 0.04	✓

Table 11. OOD Detection Performance on ImageNet Dataset. ↑ indicates higher values are better, and ↓ indicates lower values are better. We experiment with the large-scale classification on a pretrained Resnet-50 (i.e., provided by PyTorch). To avoid potential semantic or covariate overlap between ID set (ImageNet) and OOD test sets (Sun et al., 2021), we choose iNaturalist, Textures, Places365, and SUN as OOD evaluation sets following previous literatures (Liu et al., 2020; Huang et al., 2021). Here we provide results of MSP, ODIN, and Energy using FPR95 and AUROC.

ID dataset	**Method**	**OOD dataset**									
		iNaturalist		**Textures**		**Places365**		**SUN**		**Average**	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
ImageNet	MSP	47.83	89.06	49.57	85.62	61.76	84.89	61.36	85.02	55.13	85.15
	ODIN	41.39	89.83	44.15	84.04	60.12	82.46	58.52	82.66	51.29	84.75
	Energy	49.12	87.69	49.59	81.90	66.07	80.56	65.00	80.96	57.45	82.78
	MSP+UM (ours)	37.85	90.70	46.22	86.50	57.62	85.56	57.36	85.51	49.76	87.07
	ODIN+UM (ours)	28.88	91.97	**39.91**	84.89	51.22	84.12	51.06	84.01	42.77	86.25
	Energy+UM (ours)	33.11	90.91	44.10	82.78	56.52	82.53	55.44	82.64	47.29	84.72
	MSP+UMAP (ours)	36.90	91.61	51.61	87.37	61.94	86.07	61.56	84.91	53.00	87.24
	ODIN+UMAP (ours)	**21.97**	**94.71**	42.02	**88.35**	**50.06**	**86.99**	**49.69**	**86.92**	**40.94**	**89.24**
	Energy+UMAP (ours)	33.03	92.41	64.06	82.22	61.76	83.17	60.99	83.26	54.96	85.26

G. Additional Experiment Results

In this section, we provide more experiment results from different perspectives to characterize our proposed algorithms.

G.1. Additional Setups

Training details. We conduct all major experiments on DenseNet-101 (Huang et al., 2017) with training epochs fixed to 100. The models are trained using stochastic gradient descent (Kiefer & Wolfowitz, 1952) with Nesterov momentum (Duchi et al., 2011). We adopt Cosine Annealing (Loshchilov & Hutter, 2017) to schedule the learning rate which begins at 0.1. We set the momentum and weight decay to be 0.9 and 10^{-4} respectively throughout all experiments. The size of the mini-batch is 256 for both ID samples (during training and testing) and OOD samples (during testing). The choice of mask ratio for our UM and UMAP is detailed and further discussed in Appendix G.10.

Model architecture. For DenseNet-101, we fix the growth rate and reduce the rate to 12 and 0.5 respectively with the bottleneck block included in the backbone (Ming et al., 2022b). We also explore the proposed UM on WideResNet (Zagoruyko & Komodakis, 2016) with 40 depth and 4 widen factor, which is termed as WRN-40-4. The batch size for both ID and OOD testing samples is 256, and the batch size of auxiliary samples is 2000. The λ in Eq. (10) is 0.5 to keep the OE loss comparable to the CE loss. As for the outliers sampling, we randomly retrieve 50000 samples from ImageNet-1k (Deng et al., 2009) for OE and Energy (w. $\mathcal{D}_{\text{aux}}$) and 50000 samples using Thompson sampling (Thompson, 1933) for POEM (Ming et al., 2022b).

Learning rate schedules. We use 4 different learning rate schedules to demonstrate the existence of the overlaid OOD detection capability. For cosine annealing, we follow the common setups in Loshchilov & Hutter (2017); for linear schedule, the learning rate remains the same in the first one-third epochs, decreases linearly to the tenth of the initial rate in the middle one-third epochs, and decrease linearly to 1% of the initial rate in the last one-third epochs; for the multiple decay schedule, the learning rate decreases 10% of the initial rate (0.01) every 10% epochs (10 epochs); for the multiple step schedule, the learning rate decreases to 10% of the current rate every 30 epochs. All those learning rate schedules for our experiments are intuitively illustrated in Figure 8.

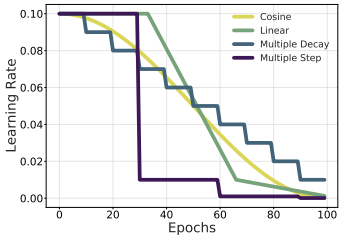


Figure 8. Learning Rate Scheduler.

G.2. Empirical verification on typical/atypical data.

In the following Tables 12, 13, 14, 15, and 16, we further conduct the experiments to identify the negative effect of learning on those atypical samples by comparing with a counterpart that learning only with the typical samples. The results demonstrate that the degeneration in detection performance is more likely to come from learning atypical samples.

In Table 12, we provide the main results for the verification using typical/atypical samples. Intuitively, we intend to separate the training dataset into a typical set and an atypical set, and train respectively on these two sets to see whether it is learning atypical samples that induce the degradation in OOD detection performance during the latter training phase. Specifically, we input the training samples through the model (DenseNet-101) of the 60th epoch and get the CE loss for selection. We provide the ACC of the generated sets on the model of the 60th epoch (ACC in the tables). The extremely low ACCs of the atypical sets show that the model of the 60th epoch can hardly predict the right label, which meets our conceptual definition of atypical samples. We then finetune the model of the 60th epoch with the generated dataset and report the OOD performance. The results show learning from only those atypical data fails to gain better detection performance than its counterpart (i.e., learning from only those typical data), although it is beneficial to improve the performance of the original multi-class classification task. The experiments provide a conceptual verification of our conjecture which links our observation and the proposed method.

G.3. Empirical Efficiency of UM and UMAP

As mentioned before, UM adopts finetuning on the proposed objective for forgetting has shown the advantages of being cost-effective compared with train-from-scratch. For the tuning epochs, we show in Figures 13 and 14 that fine-tuning using UM can converge within about 20 epochs, indicating that we can apply our UM/UMAP for far less than 100 epochs

Table 12. Fine-tuning on typical/atypical samples with different model structures (%). ↑ indicates higher values are better, and ↓ indicates lower values are better.

$\mathcal{D}_{\text{in}}$	Dataset Size	Structure	Atypical/Typical	AUROC↑	AUPR↑	FPR95↓
CIFAR-10	200	DenseNet-101	Atypical	81.45	82.40	62.10
			Typical	**82.86**	**84.38**	**60.01**
		WRN-40-4	Atypical	85.13	86.57	66.41
			Typical	**86.26**	**86.89**	**59.93**
CIFAR-100	1000	DenseNet-101	Atypical	71.96	73.16	85.57
			Typical	**74.79**	**75.83**	**80.97**
		WRN-40-4	Atypical	66.64	67.41	86.92
			Typical	**71.95**	**72.02**	**80.00**

Table 13. Fine-tuning on typical/atypical CIFAR-10 samples with DenseNet-101 (%). ↑ indicates higher values are better, and ↓ indicates lower values are better.

$\mathcal{D}_{\text{in}}$	Dataset Size	Atypical/Typical	ACC	AUROC↑	AUPR↑	FPR95↓
CIFAR-10	200	Atypical	3.50	81.45	82.40	62.10
		Typical	100.00	**82.86**	**83.48**	**60.01**
	350	Atypical	11.14	85.90	86.01	55.10
		Typical	100.00	**85.90**	**86.16**	**52.81**
	500	Atypical	16.80	84.94	85.33	59.27
		Typical	100.00	**85.53**	**86.10**	**58.74**

Table 14. Fine-tuning on typical/atypical CIFAR-10 samples with WRN-40-4 (%). ↑ indicates higher values are better, and ↓ indicates lower values are better.

$\mathcal{D}_{\text{in}}$	Dataset Size	Atypical/Typical	ACC	AUROC↑	AUPR↑	FPR95↓
CIFAR-10	200	Atypical	3.50	85.13	86.57	66.41
		Typical	100.00	**86.26**	**86.89**	**59.93**
	350	Atypical	11.14	82.92	84.24	68.57
		Typical	100.00	**85.82**	**87.84**	**65.54**
	500	Atypical	16.80	82.88	83.22	66.75
		Typical	100.00	**87.38**	**87.93**	**52.27**

Table 15. Fine-tuning on typical/atypical CIFAR-100 samples with DenseNet-101 (%). ↑ indicates higher values are better, and ↓ indicates lower values are better.

$\mathcal{D}_{\text{in}}$	Dataset Size	Atypical/Typical	ACC	AUROC↑	AUPR↑	FPR95↓
CIFAR-100	500	Atypical	1.00	72.69	73.28	80.71
		Typical	100.00	**74.07**	**75.20**	**80.19**
	800	Atypical	2.88	69.74	71.15	85.46
		Typical	100.00	**72.49**	**73.17**	**81.97**
	1000	Atypical	3.50	71.96	73.16	85.57
		Typical	100.00	**74.79**	**75.83**	**80.97**

(compared with train-from-scratch) to restore the better detection performance of the original well-trained model. It is intuitively reasonable that finetuning with the newly designed objective would benefit from the well-trained model, allowing a faster convergence since the two phases consider the same task with the same training data. As for the major experiments conducted in our work, finetuning adopts 100 epochs for better exploring and presenting its learning dynamics for research purposes, and this configuration is indicated in the training details of Section 4.1.

Table 16. Fine-tuning on typical/atypical CIFAR-100 samples with WRN-40-4 (%). ↑ indicates higher values are better, and ↓ indicates lower values are better.

$\mathcal{D}_{\text{in}}$	Dataset Size	Atypical/Typical	ACC	AUROC↑	AUPR↑	FPR95↓
CIFAR-100	500	Atypical	1.00	66.03	66.17	89.56
		Typical	100.00	**68.60**	**69.93**	**86.53**
	800	Atypical	2.88	67.59	68.66	85.61
		Typical	100.00	**70.25**	**68.95**	**79.66**
	1000	Atypical	3.50	66.64	67.41	86.92
		Typical	100.00	**71.95**	**72.02**	**80.00**

Here, we also provide an extra comparison to directly show the relative efficiency of our proposed UM/UMAP in the following Table 17 and Table 18. The results demonstrate that UM and UMAP can efficiently restore detection performance compared with the baseline. Considering the significance of the OOD awareness for those safety-critical areas, it is worthwhile to further excavate the OOD detection capability of the deployed well-trained model using our UM and UMAP.

However, there may be a concern that while both UM/UMAP and OE-based methods need extra fine-tuning processes, why should we choose UM/UMAP instead of OE-based methods, given that OE-based methods can also achieve good performance on OOD detection task. The intuition of UM/UMAP is to unleash the OOD detection capability of a pre-trained model with ID data, which is orthogonal to those OE-based methods (e.g. DOE (Wang et al., 2023)), improving the OOD capability of a pre-train model with both ID data and auxiliary data. On the one hand, OE-based methods need sampling/synthesizing large auxiliary OOD datasets, while UM/UMAP only needs the ID data. On the other hand, although both require additional costs to fine-tune the model, they are orthogonal and can be coupled (as discussed in Section 3.3). To further address the concern, we conduct additional experiments(i.e., OE-based results in Table 26) to validate their mutual benefit in the combination. According to the results, we can find that UM/UMAP with DOE achieves better performance. This is because while OE-based methods can improve the performance of OOD detection by fine-tuning with both ID and auxiliary outliers, UM/UMAP can serve as a method (only using ID data) to encourage optimization to learn a more appropriate model for OOD detection.

Table 17. Fine-tuning for 20 epochs with DenseNet-101 (%). ↑ indicates higher values are better, and ↓ indicates lower values are better.

$\mathcal{D}_{\text{in}}$	Epoch	Method	AUROC↑	AUPR↑	FPR95↓	ID-ACC↑
CIFAR-10	100	MSP	89.90	91.48	60.08	94.01
		ODIN	91.46	91.67	42.31	94.01
		Energy	92.07	92.72	42.69	94.01
		Energy + UM	93.73	94.27	33.29	92.80
		Energy + UMAP	93.97	94.38	30.71	94.01
	20	MSP + UM	90.31	91.99	53.61	91.70
		ODIN +UM	94.08	94.67	31.01	91.70
		Energy + UM	93.60	94.32	33.03	91.70
		MSP + UMAP	88.70	90.39	57.69	94.01
		ODIN + UMAP	92.88	93.33	35.19	94.01
		Energy + UMAP	92.88	93.39	35.60	94.01

G.4. Fine-grained Results on OOD Data

In order to further understand the effectiveness of the proposed UM and UMAP on different OOD datasets, we report the fine-grained results of our experiments on CIFAR-10 and CIFAR-100 with 6 OOD datasets (CIFAR-10/CIFAR-100, textures, Places365, SUN, LSUN, iNaturalist). The results on the 6 OOD datasets show the general effectiveness of the proposed UM as well as UMAP. In Table 19, **OE + UM** can outperform all the OOD baselines, and further improve the OOD performance even though the original detection performance is already well. By equipping with our proposed UM and UMAP, the baselines can outperform their counterparts on most of the OOD datasets. For instance, the FPR95 can decrease from 1.91 to 1.42. In Table 20, we also take a closer check about results on CIFAR-100 with 6 OOD datasets. Our proposed method can almost improve all competitive baselines (either the scoring functions or the finetuning with auxiliary outliers)

Table 18. Fine-tuning for 20 epochs with WRN-40-4 (%). ↑ indicates higher values are better, and ↓ indicates lower values are better.

$\mathcal{D}_{\text{in}}$	Epoch	Method	AUROC↑	AUPR↑	FPR95↓	ID-ACC↑
CIFAR-10	100	MSP	87.12	87.84	68.29	93.86
		ODIN	83.29	82.74	65.68	93.86
		Energy	87.69	88.16	58.47	93.86
		Energy + UM	91.74	92.67	40.40	92.68
		Energy + UMAP	88.84	89.31	50.23	93.86
	20	MSP + UM	89.86	91.32	51.62	91.96
		ODIN +UM	91.97	92.58	41.78	91.96
		Energy + UM	92.95	93.64	36.21	91.96
		MSP + UMAP	88.77	90.61	61.60	93.86
		ODIN + UMAP	90.85	91.89	45.70	93.86
		Energy + UMAP	91.66	92.49	42.94	93.86

Table 19. Fine-grained Results of DenseNet-101 on CIFAR-10 (%). Comparison on different OOD benchmark datasets respectively. ↑ indicates higher values are better, and ↓ indicates lower values are better.

ID dataset	**Method**	**OOD dataset**					
		CIFAR-100		**Textures**		**Places365**	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CIFAR-10	**Energy + UMAP**	54.95 ± 2.61	87.72 ± 1.05	33.59 ± 1.32	92.67 ± 0.23	32.80 ± 4.14	93.57 ± 1.12
	OE	59.29 ± 1.30	88.51 ± 0.22	2.89 ± 0.30	99.16 ± 0.05	11.14 ± 1.11	97.50 ± 0.19
	Energy (w. $\mathcal{D}_{\text{aux}}$)	79.88 ± 2.47	74.99 ± 2.40	4.27 ± 0.57	98.80 ± 0.19	14.22 ± 3.99	97.07 ± 0.72
	POEM	82.30 ± 1.57	72.74 ± 1.42	1.91 ± 0.41	99.40 ± 0.10	11.24 ± 2.70	96.67 ± 0.48
	OE + UM (ours)	$\mathbf{55.74 \pm 1.47}$	$\mathbf{89.53 \pm 0.18}$	$\mathbf{1.42 \pm 0.15}$	$\mathbf{99.49 \pm 0.04}$	$\mathbf{7.77 \pm 0.69}$	$\mathbf{98.15 \pm 0.08}$
	Energy (w. $\mathcal{D}_{\text{aux}}$)+ UM (ours)	84.52 ± 0.01	70.09 ± 0.47	8.30 ± 0.88	97.76 ± 0.05	20.27 ± 1.30	96.06 ± 0.31
	POEM + UM (ours)	84.87 ± 1.56	68.97 ± 0.39	4.73 ± 0.52	98.88 ± 0.13	19.83 ± 0.34	96.35 ± 0.09
	OE + UMAP (ours)	59.05 ± 1.41	89.14 ± 0.14	1.86 ± 0.07	99.35 ± 0.00	8.21 ± 0.12	98.07 ± 0.03
	Energy (w. $\mathcal{D}_{\text{aux}}$) + UMAP (ours)	75.18 ± 4.96	80.93 ± 4.49	2.24 ± 1.34	99.25 ± 0.29	9.30 ± 2.12	97.90 ± 0.40
	POEM + UMAP (ours)	79.33 ± 4.14	76.89 ± 4.86	2.10 ± 1.37	99.34 ± 0.30	9.94 ± 6.92	98.01 ± 1.11
	Method	**SUN**		**LSUN**		**iNaturalist**	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
	Energy + UMAP	29.05 ± 2.78	94.41 ± 0.73	2.31 ± 0.88	99.42 ± 0.04	47.22 ± 14.03	92.63 ± 2.14
	OE	8.38 ± 0.71	98.00 ± 0.14	5.90 ± 1.43	98.60 ± 0.21	5.09 ± 0.64	98.76 ± 0.11
	Energy (w. $\mathcal{D}_{\text{aux}}$)	10.30 ± 3.82	97.77 ± 0.64	12.80 ± 4.67	96.08 ± 1.38	6.93 ± 1.86	98.40 ± 0.32
	POEM	8.39 ± 2.42	98.16 ± 0.43	9.69 ± 1.89	97.25 ± 0.58	3.78 ± 0.90	98.99 ± 0.17
	OE + UM (ours)	$\mathbf{5.51 \pm 0.44}$	$\mathbf{98.55 \pm 0.07}$	$\mathbf{3.51 \pm 0.43}$	$\mathbf{98.93 \pm 0.09}$	$\mathbf{2.87 \pm 0.49}$	$\mathbf{99.14 \pm 0.09}$
	Energy (w. $\mathcal{D}_{\text{aux}}$)+ UM (ours)	16.13 ± 1.86	96.84 ± 0.30	23.27 ± 2.40	92.11 ± 0.94	11.20 ± 2.35	97.54 ± 0.42
	POEM + UM (ours)	16.16 ± 0.57	97.01 ± 0.08	25.69 ± 0.15	93.38 ± 0.27	9.30 ± 1.60	98.05 ± 0.24
	OE + UMAP (ours)	6.16 ± 0.02	98.49 ± 0.01	4.53 ± 0.16	98.86 ± 0.06	3.40 ± 0.74	98.96 ± 0.09
	Energy (w. $\mathcal{D}_{\text{aux}}$) + UMAP (ours)	6.67 ± 1.50	98.40 ± 0.32	23.50 ± 5.61	94.78 ± 2.04	3.77 ± 2.14	98.93 ± 0.46
	POEM + UMAP (ours)	7.00 ± 5.80	98.46 ± 0.96	21.17 ± 12.84	94.74 ± 3.67	3.63 ± 2.78	98.99 ± 0.54

on the 6 OOD datasets. In both w. $\mathcal{D}_{\text{aux}}$ and w.o. $\mathcal{D}_{\text{aux}}$ scenarios, Unleashing Mask can significantly excavate the intrinsic OOD detection capability of the model. In addition to unleashing the excellent OOD performance, UMAP can also maintain the high ID-ACC by learning a binary mask instead of tuning the well-trained original parameters directly. Due to the space limit, we separate the results of SVHN dataset in Tables 24 and 25 to show the relative comparison of our UM and UMAP. The results demonstrate the general effectiveness of UM/UMAP compared with the original Energy score. Besides, we find Mahalanobis performs dramatically well which is an outlier method against other post-hoc baselines when SVHN as OOD set in our experiments. Our conjecture about this phenomenon is that the Mahalanobis score can perform better on those specific OOD data by inspecting the class conditional Gaussian distributions (Lee et al., 2018a). Nonetheless, the proposed UM/UMAP can still outstrip all the baselines on most OOD datasets under various settings at the perspective of average, showing their distinguishing effectiveness and practicability.

G.5. Experiments on Different Model Structure

Following 4.2, we additionally conduct critical experiments on the WRN-40-4 (Lin et al., 2021) backbone to demonstrate the effectiveness of the proposed UM and UMAP. In Figure 11, we can find during the model training phase on ID data, there also exists the overlaid OOD detection capability can be explored in later development. In Table 21, we show the

Table 20. Fine-grained Results of DenseNet-101 on CIFAR-100 (%). Comparison on different OOD benchmark datasets respectively. ↑ indicates higher values are better, and ↓ indicates lower values are better.

ID dataset	Method	OOD dataset CIFAR-10 FPR95↓	CIFAR-10 AUROC↑	Textures FPR95↓	Textures AUROC↑	Places365 FPR95↓	Places365 AUROC↑
CIFAR-100	MSP	83.53 ± 0.33	75.11 ± 0.27	86.90 ± 0.18	71.45 ± 0.40	85.83 ± 0.48	70.54 ± 0.42
	ODIN	85.29 ± 0.17	73.31 ± 0.24	86.45 ± 1.27	71.91 ± 0.27	84.35 ± 0.64	73.58 ± 0.51
	Mahalanobis	98.25 ± 0.05	49.60 ± 1.51	$\mathbf{33.06 \pm 3.76}$	$\mathbf{90.19 \pm 1.21}$	95.20 ± 0.49	53.69 ± 1.55
	Energy	$\mathbf{82.16 \pm 0.59}$	$\mathbf{75.31 \pm 0.21}$	90.20 ± 0.30	68.98 ± 0.34	82.39 ± 0.97	73.78 ± 0.66
	Energy+UM (ours)	89.62 ± 0.07	66.12 ± 0.93	86.99 ± 1.22	65.39 ± 1.44	$\mathbf{77.30 \pm 2.08}$	$\mathbf{76.06 \pm 1.05}$
	OE	90.97 ± 0.46	69.02 ± 0.39	14.36 ± 0.25	95.92 ± 0.12	40.19 ± 6.97	90.70 ± 2.01
	Energy (w. $\mathcal{D}_{\text{aux}}$)	96.14 ± 0.06	57.52 ± 0.88	9.02 ± 0.06	97.16 ± 0.26	35.18 ± 4.73	93.29 ± 1.14
	POEM	96.19 ± 0.16	55.82 ± 1.05	7.63 ± 1.40	97.69 ± 0.06	32.67 ± 3.73	93.94 ± 0.68
	OE + UM (ours)	89.61 ± 0.08	$\mathbf{71.24 \pm 0.08}$	16.78 ± 0.25	95.60 ± 0.08	39.77 ± 0.34	91.07 ± 0.13
	Energy (w. $\mathcal{D}_{\text{aux}}$)+ UM (ours)	95.38 ± 0.45	63.41 ± 0.14	6.41 ± 0.83	97.77 ± 0.33	30.96 ± 3.61	92.85 ± 0.69
	POEM + UM (ours)	95.78 ± 0.14	60.23 ± 0.70	$\mathbf{5.17 \pm 0.18}$	$\mathbf{98.53 \pm 0.03}$	$\mathbf{23.90 \pm 0.84}$	$\mathbf{95.45 \pm 0.11}$
	OE + UMAP (ours)	$\mathbf{90.72 \pm 0.35}$	69.76 ± 0.25	15.32 ± 0.23	95.72 ± 0.01	36.42 ± 1.91	92.08 ± 0.49
	Energy (w. $\mathcal{D}_{\text{aux}}$) + UMAP (ours)	95.39 ± 0.10	63.26 ± 0.18	6.52 ± 0.44	97.83 ± 0.18	31.18 ± 0.43	93.13 ± 0.41
	POEM + UMAP (ours)	95.69 ± 0.17	61.62 ± 0.24	5.23 ± 0.58	98.52 ± 0.01	26.06 ± 1.16	94.91 ± 0.25
	Method	**SUN** FPR95↓	**SUN** AUROC↑	**LSUN** FPR95↓	**LSUN** AUROC↑	**iNaturalist** FPR95↓	**iNaturalist** AUROC↑
	MSP	88.75 ± 0.23	66.75 ± 0.25	67.83 ± 1.37	82.94 ± 0.32	85.00 ± 0.73	76.62 ± 0.25
	ODIN	88.49 ± 0.99	69.64 ± 0.61	34.80 ± 2.55	93.92 ± 0.75	81.67 ± 2.77	78.36 ± 1.57
	Mahalanobis	95.53 ± 0.37	54.37 ± 1.35	89.31 ± 4.83	43.19 ± 16.36	93.63 ± 1.19	49.60 ± 1.51
	Energy	97.17 ± 0.92	69.04 ± 0.83	35.09 ± 3.17	93.49 ± 0.87	85.70 ± 2.14	75.82 ± 1.72
	Energy+UM (ours)	$\mathbf{81.96 \pm 2.26}$	$\mathbf{71.47 \pm 1.88}$	$\mathbf{22.54 \pm 5.93}$	$\mathbf{94.98 \pm 1.75}$	$\mathbf{74.28 \pm 3.72}$	$\mathbf{80.72 \pm 3.75}$
	OE	44.47 ± 9.10	90.70 ± 2.01	$\mathbf{5.75 \pm 1.18}$	98.57 ± 0.13	25.51 ± 4.12	94.46 ± 0.88
	Energy (w. $\mathcal{D}_{\text{aux}}$)	32.69 ± 5.69	93.63 ± 1.48	55.75 ± 4.31	87.96 ± 1.03	17.34 ± 4.54	96.50 ± 0.81
	POEM	30.45 ± 5.11	94.26 ± 0.90	46.68 ± 3.59	90.30 ± 2.17	16.50 ± 2.09	96.63 ± 0.23
	OE + UM (ours)	44.23 ± 0.20	90.28 ± 0.03	5.80 ± 0.33	98.63 ± 0.03	26.72 ± 1.95	94.51 ± 0.43
	Energy (w. $\mathcal{D}_{\text{aux}}$)+ UM (ours)	28.98 ± 3.15	93.18 ± 0.69	37.56 ± 4.81	91.98 ± 0.83	10.83 ± 2.06	97.09 ± 0.69
	POEM + UM (ours)	$\mathbf{21.34 \pm 1.07}$	$\mathbf{95.76 \pm 0.22}$	33.74 ± 6.22	94.43 ± 1.18	$\mathbf{8.85 \pm 0.23}$	$\mathbf{97.93 \pm 0.03}$
	OE + UMAP (ours)	39.58 ± 2.02	91.37 ± 0.52	5.77 ± 0.71	$\mathbf{98.64 \pm 0.12}$	23.33 ± 1.24	95.08 ± 0.22
	Energy (w. $\mathcal{D}_{\text{aux}}$) + UMAP (ours)	29.65 ± 1.06	93.38 ± 0.19	35.94 ± 0.75	92.08 ± 0.39	13.96 ± 2.48	96.87 ± 0.15
	POEM + UMAP (ours)	23.73 ± 0.71	95.25 ± 0.06	33.09 ± 5.94	93.57 ± 0.88	9.76 ± 1.09	97.77 ± 0.24

Table 21. Results of WRN-40-4. Comparison with competitive OOD detection baselines (%). We respectively train WRN-40-4 on CIFAR-10 and CIFAR-100. For those methods involving outliers, we retrieve 5000 samples from ImageNet-1k. ↑ indicates higher values are better, and ↓ indicates lower values are better.

$\mathcal{D}_{\text{in}}$	Method	AUROC↑	AUPR↑	FPR95↓	ID-ACC↑
CIFAR-10	MSP(Hendrycks & Gimpel, 2017)	87.12 ± 0.25	87.84 ± 0.30	68.29 ± 0.96	$\mathbf{93.86 \pm 0.19}$
	ODIN(Liang et al., 2018)	83.29 ± 0.72	82.74 ± 0.79	65.68 ± 0.77	$\mathbf{93.86 \pm 0.19}$
	Mahalanobis(Lee et al., 2018b)	77.57 ± 0.28	76.11 ± 0.10	61.18 ± 0.10	$\mathbf{93.86 \pm 0.19}$
	Energy(Liu et al., 2020)	87.69 ± 0.54	88.16 ± 0.69	58.47 ± 1.94	$\mathbf{93.86 \pm 0.19}$
	Energy+UM (ours)	$\mathbf{91.74 \pm 0.43}$	$\mathbf{92.67 \pm 0.52}$	$\mathbf{40.40 \pm 1.32}$	92.68 ± 0.23
	Energy+UMAP (ours)	88.84 ± 1.02	89.31 ± 1.44	50.23 ± 2.25	$\mathbf{93.86 \pm 0.19}$
CIFAR-100	MSP(Hendrycks & Gimpel, 2017)	72.34 ± 0.63	72.69 ± 0.44	85.40 ± 0.59	$\mathbf{75.01 \pm 0.07}$
	ODIN(Liang et al., 2018)	68.78 ± 0.67	66.92 ± 0.72	85.28 ± 0.64	$\mathbf{75.01 \pm 0.07}$
	Mahalanobis(Lee et al., 2018b)	68.20 ± 0.99	68.30 ± 1.15	76.46 ± 2.02	$\mathbf{75.01 \pm 0.07}$
	Energy(Liu et al., 2020)	74.00 ± 0.41	73.02 ± 0.47	81.37 ± 0.08	$\mathbf{75.01 \pm 0.07}$
	Energy+UM (ours)	76.07 ± 0.04	76.94 ± 0.06	74.29 ± 1.66	59.08 ± 2.75
	Energy+UMAP (ours)	$\mathbf{77.35 \pm 0.78}$	$\mathbf{77.43 \pm 0.91}$	$\mathbf{68.20 \pm 0.06}$	$\mathbf{75.01 \pm 0.07}$

comparison of multiple OOD detection baselines, evaluating the OOD performance on the different OOD datasets mentioned in Section 4.1. The results again demonstrate that our proposed method indeed excavates the intrinsic detection capability and improves the performance.

As for the fine-grained results of WRN-40-4, we report results on 6 OOD datasets respectively. When trained on CIFAR-10, UM can outstrip all the scoring function baselines on 5 OOD datasets except Textures on which Mahalanobis performs better while UMAP still has excellent OOD performance ranking only second to UM. When trained on CIFAR-100, UM and

Table 22. Fine-grained Results of WRN-40-4 on CIFAR-10 (%). Comparison on different OOD benchmark datasets. ↑ indicates higher values are better, and ↓ indicates lower values are better.

ID dataset	Method	OOD dataset					
		CIFAR-100		Textures		Places365	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CIFAR-10	MSP	70.96 ± 0.70	86.08 ± 0.08	68.81 ± 1.29	86.53 ± 0.83	68.31 ± 0.25	86.71 ± 0.13
	ODIN	64.97 ± 0.08	83.36 ± 0.11	66.86 ± 2.24	81.34 ± 0.81	66.49 ± 1.16	83.47 ± 0.93
	Mahalanobis	79.84 ± 0.55	70.33 ± 0.24	$\mathbf{22.56 \pm 0.08}$	$\mathbf{94.07 \pm 0.04}$	85.09 ± 0.59	67.90 ± 0.37
	Energy	61.09 ± 0.58	86.66 ± 0.04	64.29 ± 1.72	85.56 ± 0.53	55.32 ± 0.13	88.29 ± 0.26
	Energy+UM (ours)	$\mathbf{57.21 \pm 1.41}$	$\mathbf{87.56 \pm 0.15}$	46.49 ± 1.03	89.74 ± 0.45	$\mathbf{40.68 \pm 4.46}$	$\mathbf{92.51 \pm 0.97}$
	Energy+UMAP (ours)	65.45 ± 1.10	84.65 ± 0.95	59.14 ± 1.64	85.27 ± 1.74	48.16 ± 1.89	90.43 ± 0.47
	Method	**SUN**		**LSUN**		**iNaturalist**	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
	MSP	68.62 ± 0.50	86.95 ± 0.23	52.97 ± 3.07	92.41 ± 0.18	76.05 ± 0.01	83.44 ± 0.36
	ODIN	65.47 ± 0.78	83.79 ± 1.10	31.89 ± 3.44	94.34 ± 0.89	79.28 ± 0.18	79.80 ± 0.35
	Mahalanobis	82.92 ± 0.28	70.52 ± 0.47	64.31 ± 0.57	67.75 ± 0.55	81.50 ± 2.91	74.97 ± 2.91
	Energy	54.88 ± 0.18	88.67 ± 0.30	24.99 ± 1.38	95.98 ± 0.37	75.89 ± 0.85	82.40 ± 0.22
	Energy+UM (ours)	$\mathbf{38.92 \pm 3.46}$	$\mathbf{92.98 \pm 0.95}$	$\mathbf{8.38 \pm 0.77}$	$\mathbf{98.18 \pm 0.16}$	$\mathbf{66.02 \pm 6.70}$	$\mathbf{85.22 \pm 3.42}$
	Energy+UMAP (ours)	45.94 ± 2.64	91.27 ± 0.56	14.10 ± 0.04	97.46 ± 0.14	74.69 ± 0.15	81.13 ± 1.12

Table 23. Fine-grained Results of WRN-40-4 on CIFAR-100 (%). Comparison on different OOD benchmark datasets. ↑ indicates higher values are better, and ↓ indicates lower values are better.

ID dataset	Method	OOD dataset					
		CIFAR-10		Textures		Places365	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
CIFAR-100	MSP	83.83 ± 0.29	75.50 ± 0.21	86.15 ± 0.23	72.36 ± 0.40	86.72 ± 0.29	69.60 ± 0.27
	ODIN	83.70 ± 0.30	74.32 ± 0.03	81.57 ± 1.74	71.67 ± 0.28	88.07 ± 0.11	64.83 ± 1.36
	Mahalanobis	96.89 ± 0.11	68.78 ± 0.67	$\mathbf{31.02 \pm 2.04}$	$\mathbf{91.85 \pm 0.91}$	93.34 ± 0.44	61.28 ± 1.62
	Energy	$\mathbf{81.32 \pm 0.47}$	$\mathbf{77.49 \pm 0.26}$	86.38 ± 0.49	73.50 ± 0.45	84.45 ± 0.38	69.82 ± 0.63
	Energy+UM (ours)	89.23 ± 1.51	63.85 ± 1.73	78.90 ± 0.07	72.58 ± 1.33	$\mathbf{80.46 \pm 1.99}$	70.49 ± 1.01
	Energy+UMAP (ours)	94.11 ± 0.72	60.77 ± 0.96	66.94 ± 4.49	75.82 ± 5.33	82.59 ± 0.27	$\mathbf{71.92 \pm 3.58}$
	Method	**SUN**		**LSUN**		**iNaturalist**	
		FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑
	MSP	88.88 ± 0.83	65.22 ± 0.85	78.56 ± 0.66	79.10 ± 0.53	86.72 ± 0.29	73.75 ± 0.56
	ODIN	91.00 ± 0.10	59.06 ± 1.81	70.14 ± 2.42	84.03 ± 0.86	87.86 ± 1.13	64.52 ± 1.51
	Mahalanobis	94.22 ± 0.01	60.09 ± 1.56	89.73 ± 2.87	40.81 ± 3.07	87.25 ± 3.28	74.98 ± 2.85
	Energy	88.35 ± 0.52	64.04 ± 0.76	59.84 ± 0.06	87.91 ± 0.53	88.91 ± 0.78	67.81 ± 0.91
	Energy+UM (ours)	84.04 ± 0.09	67.19 ± 0.14	33.87 ± 1.21	92.29 ± 0.02	76.91 ± 6.07	79.28 ± 4.17
	Energy+UMAP (ours)	$\mathbf{80.53 \pm 1.31}$	$\mathbf{72.68 \pm 4.50}$	$\mathbf{27.79 \pm 2.19}$	$\mathbf{93.39 \pm 0.57}$	$\mathbf{55.53 \pm 4.61}$	$\mathbf{85.65 \pm 0.83}$

UMAP can also outperform the baselines on most OOD datasets. The fine-grained results of WRN-40-4 further demonstrate the effectiveness of the proposed UM/UMAP on other architectures. The future extension can also take other advanced model structures for OOD detection (Ming et al., 2022a) into consideration.

G.6. Additional Experiments on More Advanced Post-hoc and OE-based Methods

Except for some representative methods (like MSP, Energy, OE, POEM) that have been considered in the experiments, in Table 26, we add more advanced post-hoc and OE-based methods (Sun et al., 2021; Sun & Li, 2022; Djurisic et al., 2023; Katz-Samuels et al., 2022; Wang et al., 2023) as comparison to further validate the effectiveness of the proposed UM/UMAP.

G.7. Additional Verification for Intrinsic OOD Discriminative Capability

In Section 4.3, we display the overlaid OOD detection capability on CIFAR-10 using SVHN as the OOD dataset. Here, we additionally verify the previously observed trend during training when training DenseNet-101 on CIFAR-100 using iNaturalist as an OOD dataset. In Figure 9, we trace the three evaluation metrics during training on CIFAR-100 using 4 different learning rate schedules. Consistent with the original experiment, we still use iNaturalist as the OOD dataset. It can be seen for all three metrics that exists a middle stage where the model has the better OOD detection capability (For FPR95,

Table 24. Results of DenseNet-101 when SVHN as OOD set (%). Comparison on different ID benchmark datasets. ↑ indicates higher values are better, and ↓ indicates lower values are better.

OOD dataset	Method	ID dataset					
		CIFAR-10			CIFAR-100		
		AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓
SVHN	MSP	90.67 ± 1.35	87.36 ± 2.53	63.35 ± 3.54	75.00 ± 4.09	64.57 ± 5.21	84.11 ± 4.64
	ODIN	92.19 ± 1.14	86.30 ± 2.17	43.27 ± 5.07	72.64 ± 2.02	60.22 ± 2.09	91.43 ± 2.32
	Mahalanobis	$\mathbf{98.12 \pm 0.92}$	93.76 ± 5.52	$\mathbf{7.09 \pm 1.95}$	$\mathbf{92.63 \pm 1.04}$	$\mathbf{82.93 \pm 5.01}$	$\mathbf{32.55 \pm 2.93}$
	Energy	92.84 ± 1.86	89.23 ± 3.62	46.37 ± 7.31	77.64 ± 3.76	68.45 ± 5.28	86.50 ± 2.98
	Energy+UM (ours)	95.42 ± 1.77	93.29 ± 2.61	28.90 ± 13.53	78.80 ± 7.41	70.60 ± 9.59	85.67 ± 7.28
	Energy+UMAP (ours)	97.37 ± 1.76	$\mathbf{95.67 \pm 2.45}$	15.04 ± 11.34	83.31 ± 5.31	75.10 ± 6.76	75.25 ± 9.74
	OE	98.97 ± 0.05	98.15 ± 0.15	3.91 ± 0.16	95.50 ± 1.49	91.86 ± 2.52	21.67 ± 8.92
	Energy (w. $\mathcal{D}_{\text{aux}}$)	98.93 ± 0.23	98.07 ± 0.36	3.17 ± 0.33	96.38 ± 0.63	94.13 ± 1.01	19.14 ± 6.80
	POEM	94.37 ± 0.07	94.50 ± 0.06	18.50 ± 0.33	94.00 ± 0.98	89.25 ± 1.99	36.56 ± 4.43
	OE + UM (ours)	99.43 ± 0.09	$\mathbf{98.98 \pm 0.17}$	$\mathbf{1.73 \pm 0.41}$	95.96 ± 0.95	92.76 ± 1.53	20.09 ± 6.58
	Energy (w. $\mathcal{D}_{\text{aux}}$)+ UM (ours)	98.16 ± 0.48	96.67 ± 0.95	7.18 ± 2.74	96.22 ± 1.78	93.40 ± 2.75	18.46 ± 12.78
	POEM + UM (ours)	98.63 ± 0.32	97.49 ± 0.51	4.08 ± 1.41	95.90 ± 0.25	93.89 ± 0.01	26.66 ± 4.39
	OE + UMAP (ours)	$\mathbf{99.46 \pm 0.01}$	98.85 ± 0.01	2.27 ± 0.00	95.05 ± 0.28	90.57 ± 0.18	24.20 ± 2.05
	Energy (w. $\mathcal{D}_{\text{aux}}$) + UMAP (ours)	99.21 ± 0.11	98.73 ± 0.21	1.91 ± 0.47	$\mathbf{97.05 \pm 0.99}$	$\mathbf{94.41 \pm 1.99}$	$\mathbf{12.52 \pm 4.99}$
	POEM + UMAP (ours)	99.36 ± 0.19	98.73 ± 0.47	2.58 ± 1.53	96.04 ± 0.95	93.83 ± 1.36	23.57 ± 5.85

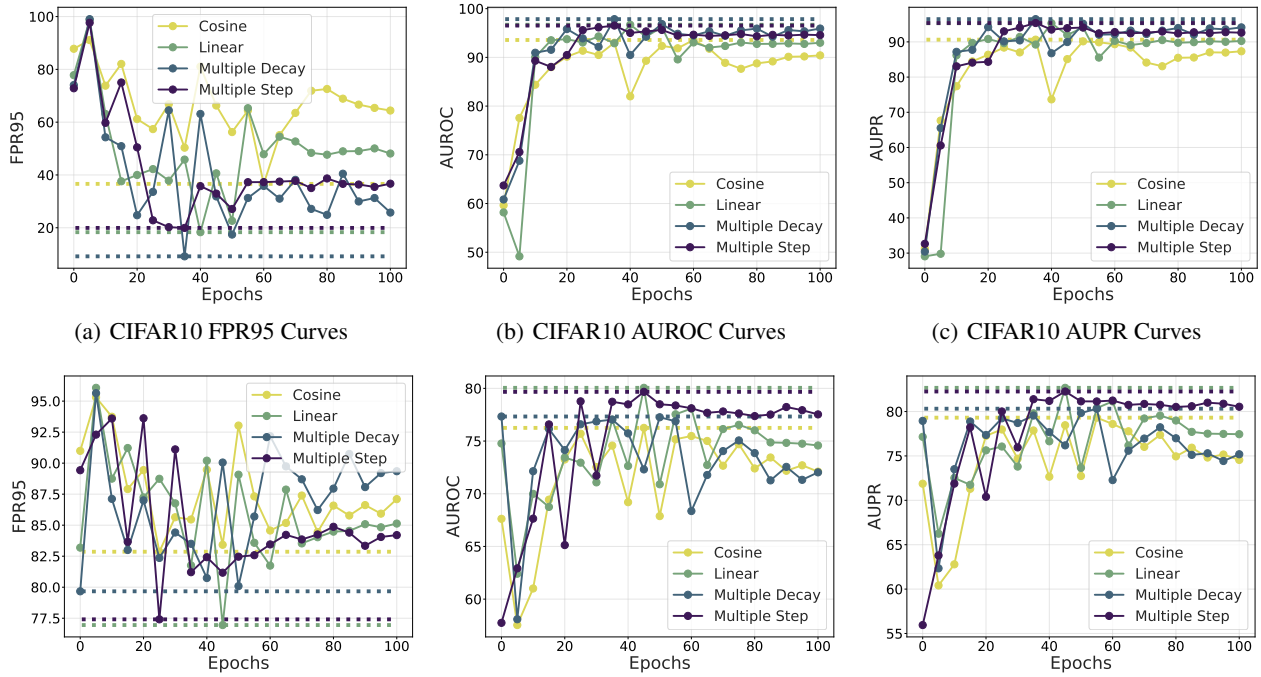
Table 25. Results of WRN-40-4 when SVHN as OOD set (%). Comparison on different ID benchmark datasets. ↑ indicates higher values are better, and ↓ indicates lower values are better.

OOD dataset	Method	ID dataset					
		CIFAR-10			CIFAR-100		
		AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓
SVHN	MSP	87.68 ± 0.65	81.94 ± 1.75	72.35 ± 0.88	70.83 ± 2.06	56.20 ± 0.88	86.30 ± 2.09
	ODIN	76.98 ± 6.71	65.19 ± 8.60	84.78 ± 8.44	63.00 ± 0.59	47.27 ± 0.18	94.63 ± 1.08
	Mahalanobis	$\mathbf{97.48 \pm 0.22}$	$\mathbf{94.49 \pm 0.80}$	$\mathbf{12.07 \pm 0.71}$	$\mathbf{91.26 \pm 2.35}$	$\mathbf{84.19 \pm 3.42}$	$\mathbf{42.72 \pm 11.32}$
	Energy	86.26 ± 3.92	79.81 ± 5.42	72.80 ± 13.80	77.47 ± 1.78	64.99 ± 1.57	80.36 ± 1.70
	Energy+UM (ours)	95.96 ± 1.07	93.79 ± 1.42	25.09 ± 9.32	86.77 ± 2.93	82.39 ± 2.40	76.59 ± 13.00
	Energy+UMAP (ours)	91.67 ± 4.21	86.15 ± 6.36	44.16 ± 17.68	81.24 ± 5.09	69.53 ± 7.29	69.90 ± 12.66

Table 26. Results of additional comparison with advanced methods of post-hoc scoring functions or fine-tuning with auxiliary outliers (%). ↑ indicates higher values are better, and ↓ indicates lower values are better.

$\mathcal{D}_{\text{in}}$	Method	AUROC↑	AUPR↑	FPR95↓	ID-ACC↑	w./w.o $\mathcal{D}_{\text{aux}}$
CIFAR-10	ReAct(Sun et al., 2021)	92.76 ± 0.26	93.57 ± 0.34	38.43 ± 1.31	$\mathbf{93.74 \pm 0.10}$	
	DICE(Sun & Li, 2022)	90.66 ± 0.94	91.11 ± 1.04	41.51 ± 2.39	92.55 ± 0.20	
	ASH-S(Djurisic et al., 2023)	95.13 ± 0.08	95.51 ± 0.14	25.87 ± 0.22	93.69 ± 0.10	
	ASH-S+UM(ours)	$\mathbf{95.22 \pm 0.27}$	$\mathbf{95.53 \pm 0.26}$	$\mathbf{23.94 \pm 1.27}$	92.47 ± 0.13	
	ASH-S+UMAP(ours)	94.57 ± 0.94	94.86 ± 0.95	25.95 ± 3.24	93.69 ± 0.10	
	WOODS(Katz-Samuels et al., 2022)	97.32 ± 0.04	96.78 ± 0.12	12.94 ± 0.46	$\mathbf{92.22 \pm 0.84}$	✓
	DOE(Wang et al., 2023)	97.41 ± 0.01	97.38 ± 0.04	12.84 ± 0.23	91.02 ± 0.02	✓
	DOE+UM(ours)	$\mathbf{97.49 \pm 0.00}$	$\mathbf{97.63 \pm 0.03}$	$\mathbf{11.62 \pm 0.11}$	90.57 ± 0.32	✓
	DOE+UMAP(ours)	97.48 ± 0.01	96.62 ± 0.05	11.85 ± 0.16	91.02 ± 0.02	✓
CIFAR-100	ReAct(Sun et al., 2021)	75.66 ± 0.30	75.49 ± 0.24	79.04 ± 0.39	73.69 ± 0.22	
	DICE(Sun & Li, 2022)	78.41 ± 0.65	79.11 ± 0.52	69.33 ± 2.35	67.77 ± 0.47	
	ASH-S(Djurisic et al., 2023)	83.53 ± 0.07	84.02 ± 0.10	61.78 ± 0.41	$\mathbf{74.09 \pm 0.13}$	
	ASH-S+UM(ours)	$\mathbf{83.60 \pm 0.37}$	$\mathbf{84.27 \pm 0.55}$	$\mathbf{56.56 \pm 2.13}$	64.39 ± 0.87	
	ASH-S+UMAP(ours)	81.98 ± 0.04	82.84 ± 0.01	59.10 ± 0.01	$\mathbf{74.09 \pm 0.13}$	
	WOODS(Katz-Samuels et al., 2022)	92.68 ± 0.12	87.18 ± 0.79	33.10 ± 1.85	$\mathbf{73.08 \pm 0.97}$	✓
	DOE(Wang et al., 2023)	92.77 ± 0.01	86.79 ± 0.00	30.63 ± 0.11	73.01 ± 0.17	✓
	DOE+UM(ours)	92.93 ± 0.08	87.33 ± 0.24	$\mathbf{29.47 \pm 0.15}$	71.11 ± 0.03	✓
	DOE+UMAP(ours)	$\mathbf{92.95 \pm 0.01}$	$\mathbf{87.41 \pm 0.05}$	29.61 ± 0.32	73.01 ± 0.17	✓

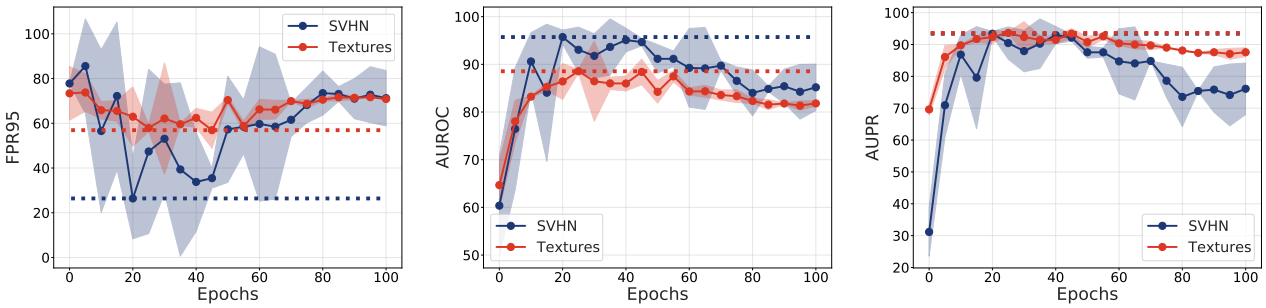
it is smaller (better) in the middle stage; for AUROC and AUPR, they are higher (better) in the middle stage). Besides that, we also look into the change of OOD performance on other architecture (e.g., WRN-40-4) in Figure 10 and Figure 11. In



(a) CIFAR10 FPR95 Curves (b) CIFAR10 AUROC Curves (c) CIFAR10 AUPR Curves

(d) CIFAR100 FPR95 Curves (e) CIFAR100 AUROC Curves (f) CIFAR100 AUPR Curves

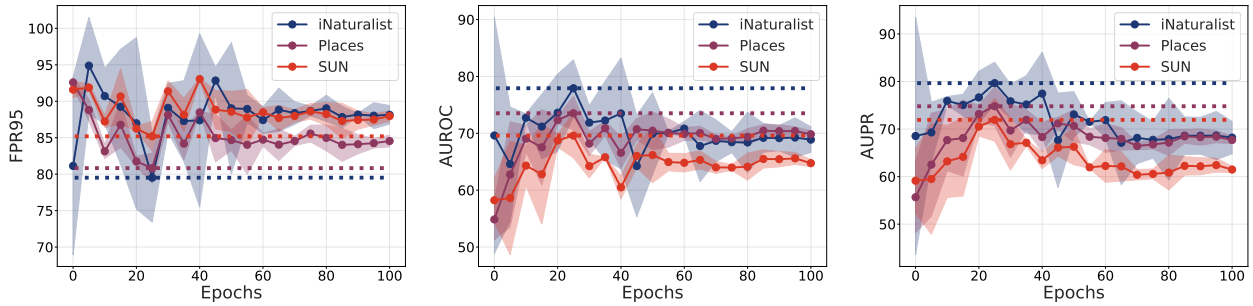
Figure 9. Ablation studies on three metrics with 4 different learning rate schedules. The model is DenseNet-101 trained on CIFAR-100 with iNaturalist as the OOD dataset. (a) change of FPR95 throughout the pruning phase when training on CIFAR-100; (b) change of AUROC throughout the pruning phase when training on CIFAR-100; (c) change of AUPR throughout the pruning phase when training on CIFAR-100. It demonstrates a better middle stage exists according to the three metrics.



(a) FPR95 Curves (b) AUROC Curves (c) AUPR Curves

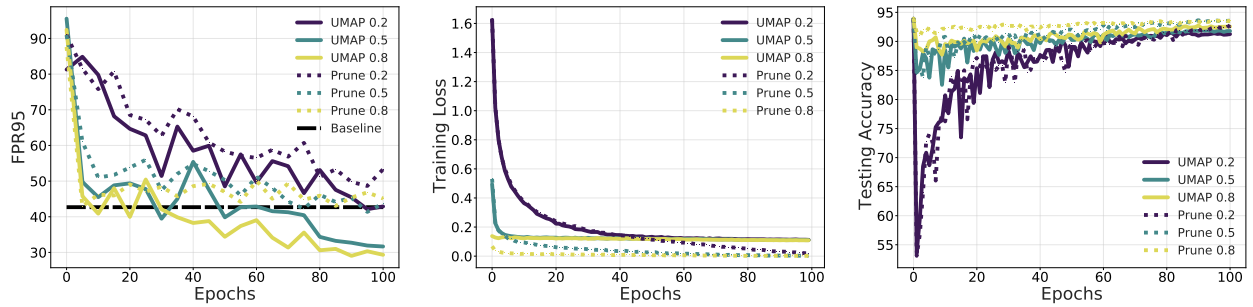
Figure 10. Ablation studies on three metrics of WRN-40-4 with CIFAR-10 as ID dataset, SVHN, and Textures as OOD datasets. (a) change of FPR95 throughout the pruning phase when training on CIFAR-10; (b) change of AUROC throughout the pruning phase when training on CIFAR-10; (c) change of AUPR throughout the pruning phase when training on CIFAR-10. It demonstrates a better middle stage exists according to the three metrics.

Figure 10, we display the curves of three metrics of WRN-40-4 when trained on CIFAR-10 with SVHN and Textures as OOD datasets. The trend that the OOD performance first goes better and then converges to worse OOD performance can be reflected. In Figure 11, we continually provide curves of the three metrics of WRN-40-4 during training on CIFAR-100 with iNaturalist, Places365, and SUN as OOD datasets. A clear better middle stage can still be excavated in this scenario.



(a) FPR95 Curves (b) AUROC Curves (c) AUPR Curves

Figure 11. Ablation studies on three metrics of WRN-40-4 with CIFAR-100 as ID dataset, iNaturalist, Places365, and SUN as OOD datasets. (a) change of FPR95 throughout the pruning phase when training on CIFAR-100; (b) change of AUROC throughout the pruning phase when training on CIFAR-100; (c) change of AUPR throughout the pruning phase when training on CIFAR-100. It demonstrates a better middle stage exists according to the three metrics.



(a) FPR95 Curves (b) Train Loss Curves (c) Test Acc Curves

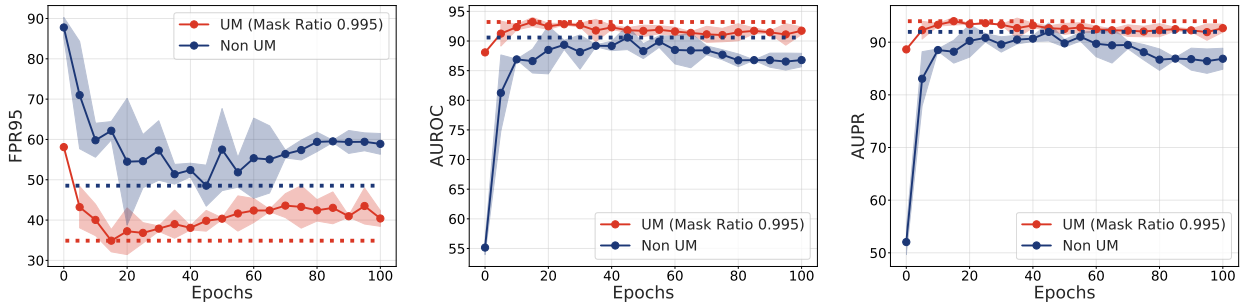
Figure 12. Ablation studies on Prune Rate of UMAP. (a) change of OOD performance throughout the pruning phase; (b) training loss converges to estimated loss constraint properly; (c) though ID-ACC is not taken into consideration for UMAP, it still rises high after training for 100 epochs.

G.8. Ablation on UMAP which Adopting Pruning on UM.

We conduct various experiments to see whether pruning has an impact on Unleashing Mask itself. To be specific, we expect the pruning to learn a mask on the given model while not impairing the excellent OOD performance that UM brings. In Figure 12, it presents that pruning from a wide range (e.g. $p \in [0.3, 0.9]$) can well maintain the effectiveness of UM while possessing a terrific convergence trend. For simplicity, we use prune to indicate the original pruning approach and UMAP indicate UM with pruning on the mask with our newly designed forgetting objective in Figure 12. In Figure 12(a), the solid lines represent the proposed UMAP and the dashed lines represent only pruning the well-trained model at prune rates 0.2, 0.5, and 0.8. While the model's OOD performance can't be improved (not better than the baseline) through only pruning, using our proposed forgetting objective for the loss constrain can significantly bring out better OOD performance at a wide range of mask rates (e.g. $p \in [0.5, 0.8]$). In Figure 12(b), we intuitively reflect the effect of the estimated loss constraint by the initialized mask which redirects the gradients when the loss reaches the value, while the loss will just approach 0 when pruning only. In Figure 12(c), we can see that ID-ACC for both UMAP and Prune can converge to approximately the same high level ($92\% \sim 94\%$), though we can simply remove the learned mask to recover the original ID-ACC.

G.9. Fine-grained comparison of model weights.

We display the weights of the original model, pruned model, and the UMAP model respectively in Figure 15. The histograms show that the adopted pruning algorithm tends to choose weights far from 0 for the first convolution layer, shown in Figure 15(a). However, for almost all layers (from the 2nd to the 98th), the pruning chooses weights with no respect to the



(a) FPR95 Curves (b) AUROC Curves (c) AUPR Curves

Figure 13. Ablation studies to reflect the effectiveness of UM. The mask ratio of UM is 99.5%. (a) change of FPR95 throughout the training phase on CIFAR-10; (b) change of AUROC throughout the training phase on CIFAR-10; (c) change of AUPR throughout the training phase on CIFAR-10.

value of weights, shown in Figure 15(b). For the fully connected layer, the pruning algorithm itself still keeps its behavior on the first layer, while UMAP forces the pruning algorithm to choose weights near 0, shown in Figure 15(c), indicating that forgetting learned atypical samples doesn't necessarily correspond to larger weights or smaller weights.

G.10. The effectiveness of UM

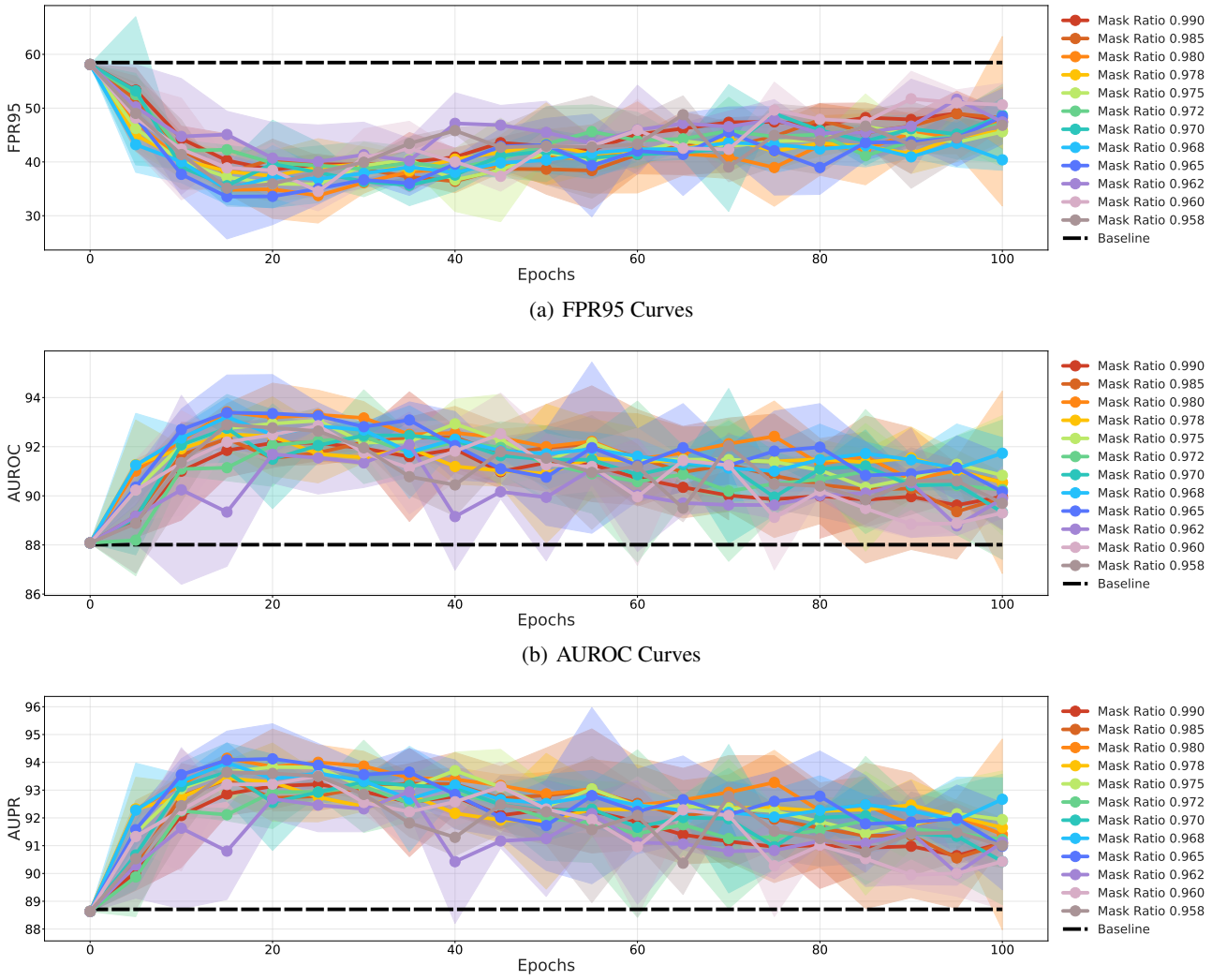
In Figure 13, we present the FPR95, AUROC, and AUPR curves during training to show the comparison of the original training and our proposed UM on ID data. We observe that training using UM can consistently outperform the vanilla model training, either for the final stage or the middle stage with the best OOD detection performance indicated by the FPR95 curve. In Figure 14, we also adopt different mask rates for the initialized loss constraint estimation for forgetting the atypical samples. The results show that a wide range of mask ratios (i.e., from 96% to 99%) to estimate the loss constraint used in Eq. (3) can gain better OOD detection performance than the baseline. It shows the mask ratio would be robust to hyper-parameter selection under a certain small value. The principle intuition behind this is our revealed important observation as indicated in Figures 1(a), 2(a), and 2(b). With the guidance of the general mechanism, empirically choosing the hyper-parameter using the validation set is supportable and valuable for excavating better OOD detection capability of the model as conducted by previous literature (Hendrycks et al., 2019b; Liu et al., 2020; Sun et al., 2021).

In our experiments, we empirically determine the value of our proposed UM and UMAP by examining the training loss on the masked output. For CIFAR-10 as ID datasets, the value of the mask ratio is 97.5%, and the estimated loss constraint for forgetting is 0.10 for our tuning until the convergence; For CIFAR-100, the value of the mask ratio is 97%, and the estimated loss constraint for forgetting is 1.20 for our tuning until the convergence. To choose the parameters of the estimated loss constraint, we use the TinyImageNet (Tavanaei, 2020) dataset as the validation set, which is not seen during training and is not considered in our evaluation of OOD detection performance. Since the core intuition behind our method is to restore the OOD detection performance starting from the well-trained model stage, forgetting a relatively small portion (empirically found around 97% mask ratio) of atypical samples can be beneficial for the two common benchmarked datasets. In addition, we also verify the effectiveness of UM and UMAP considering the large-scale ImageNet as ID dataset in Table 11 and Appendix F, the loss constraint for forgetting can be 0.6 which is estimated using the mask ratio as 99.6%. To find the optimal parameter for tuning, more advanced searching techniques like AutoML or validation design based on the important observation in our work may be further employed in the future. For the safety concerns, it is affordable and reasonable to gain significant OOD detection performance improvement by investing extra computing resources.

H. Summarization of the proposed UM/UMAP's advantages

Regarding the advantages of the proposed method, we kindly interpret them as follows,

- **Novelty.** The proposed UM/UMAP is the first to emphasize the intrinsic OOD detection capability of a given well-trained model during its training phase, better leveraging what has been learned and drawing new insight into the relationship between the OOD detection and the original classification task. This work also shows that ID data is



(a) FPR95 Curves

(b) AUROC Curves

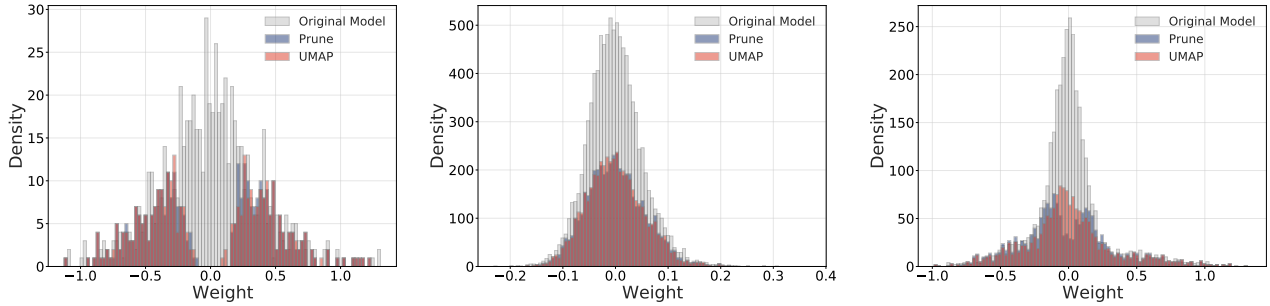
(c) AUPR Curves

Figure 14. Ablation studies of WRN-40-4 on various Mask Ratios. The mask rate is from 95.8% to 99.0%. (a) change of FPR95 throughout the training on CIFAR-10; (b) change of AUROC throughout the training on CIFAR-10; (c) change of AUPR throughout the training on CIFAR-10.

important for a well-trained model's OOD discriminative capability.

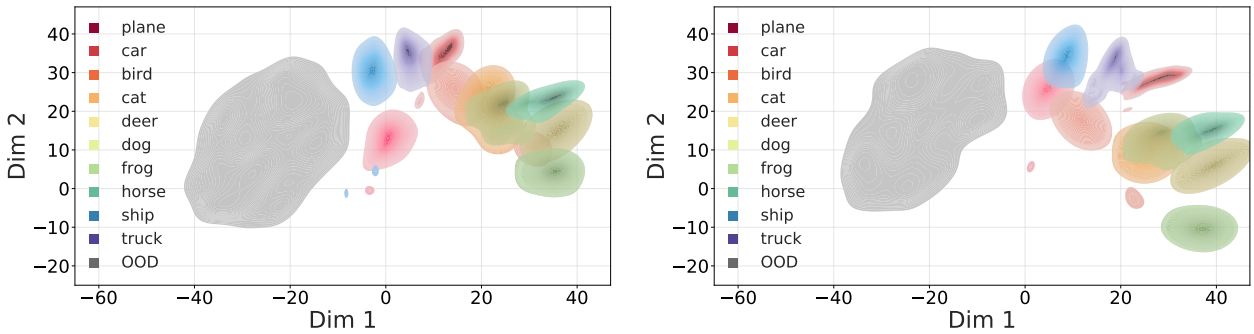
- **Simplicity.** Based on the empirical insights that atypical semantics may impair the OOD detection capability, we introduced the easy-to-adopt forgetting objective to weaken the influence of atypical samples on the OOD detection performance. Besides, to maintain the ID performance, we proposed to learning a mask instead of tuning the model directly. Such a design makes UM/UMAP easy to follow and a good starting point to conduct further adjustments. They build on extensive empirical analysis on the point of how to unleash the optimal OOD detection capacity of one given model. Besides, this work explores an orthogonal perspective to previous methods, which shows the consistent improvement combined with previous methods in a range of experiments.

- **Compatibility & Effectiveness.** UM/UMAP is orthogonal to other competitive methods and can be flexibly combined with them. Extensive experiments demonstrate that UM/UMAP can consistently improve the baselines on average in both benchmarked datasets and large-scale ImageNet (e.g., Tables 1,2,10,11,19,20,21,22,23,24,25; Figures 4(a),4(d),4(e),12,13,14,16,17(c),17(d).



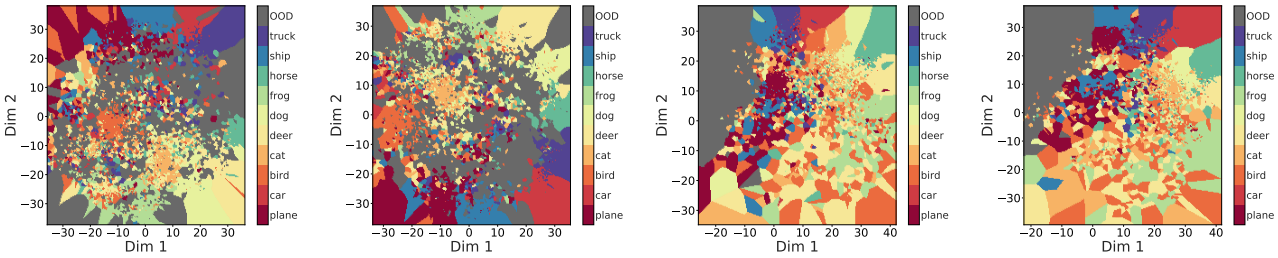
(a) First Layer (b) 50th Layer (c) Last Layer (FC)

Figure 15. Histograms of different layers for the original model, pruned model, and UMAP model. The model is DenseNet-101 with a prune rate of 50%. (a) the histogram of the first convolution layer; (b) the histogram of the 50th convolution layer; (c) the histogram of the last (fully connected) layer.



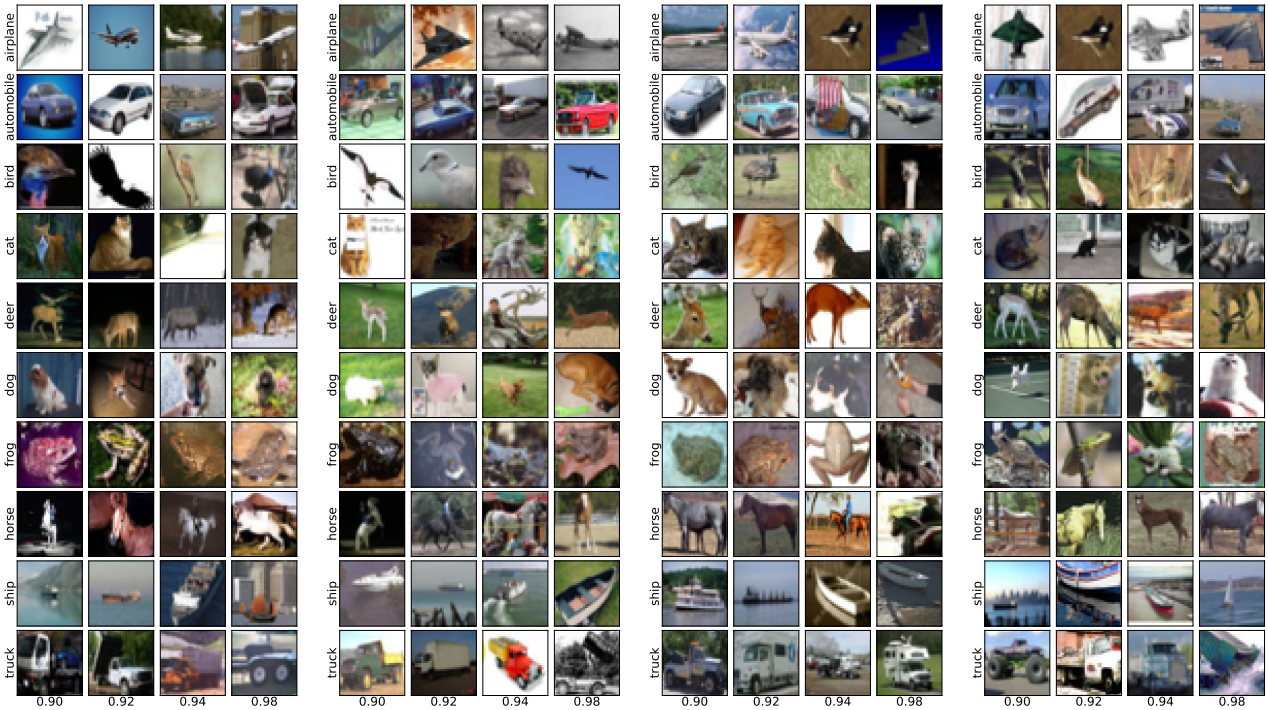
(a) UM (b) UMAP

Figure 16. Similar to previous Figure 2(e), we visualize the learned feature by UM and UMAP: (a) TSNE visualization of UM; (b) TSNE visualization of UMAP. Compared to the visualization in Figure 2(e), it is apparent that ID distribution and OOD distribution have larger intervals, which indicates that the UM/UMAP-processed model can better distinguish ID and OOD samples. Moreover, the OOD distributions are more united compared to those in Figure 2(e), indicating better OOD discriminative Capability.



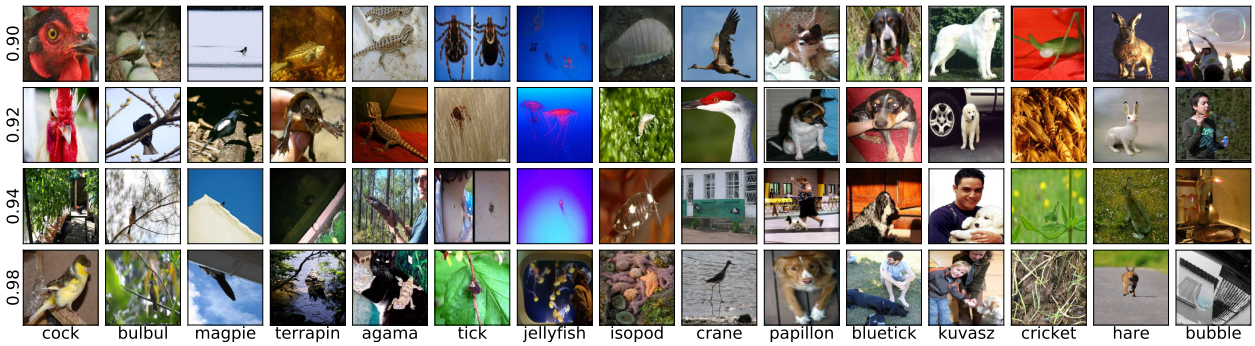
(a) Epoch 60 (b) Epoch 100 (c) UM (d) UMAP

Figure 17. The visualization of the decision boundary: (a) decision boundary of epoch 60; (b) decision boundary of epoch 100; (c) decision boundary of UM; (d) decision boundary of UMAP. Based on Figures 2(e) and 16, we further use the TSNE embeddings to simulate the decision boundaries generated with K-NN (Fix & Hodges, 1989). The visualization shows that UM/UMAP can help exclude the OOD distribution from the ID distribution, making it easier for the model to distinguish ID/OOD data. While the OOD performance at epoch 60 is better than that at epoch 100, the OOD distribution gets mixed up with the ID distribution in both figures. Though the simulation can't act as a concrete reflection of the model's feature space, it still can intuitively explain the mechanism of UM/UMAP.

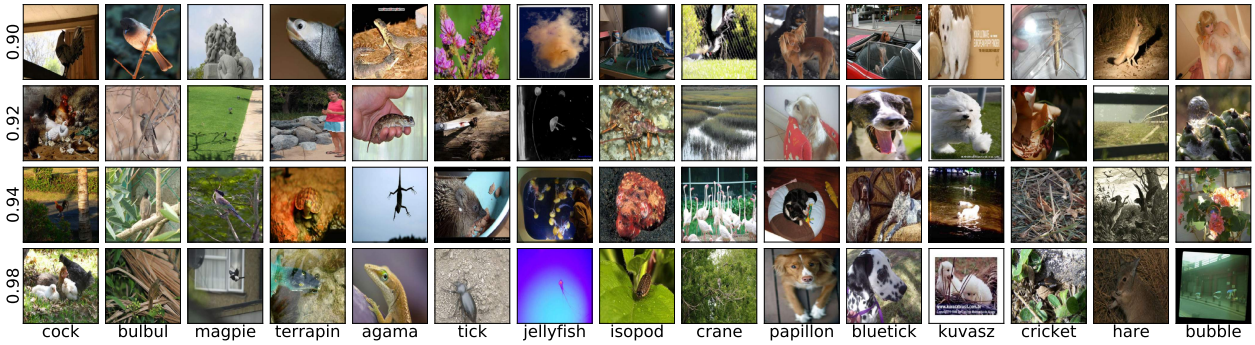


(a) Layer-wise masking scores (b) Layer-wise masking weights (c) Model-wise masking scores (d) Model-wise masking weights

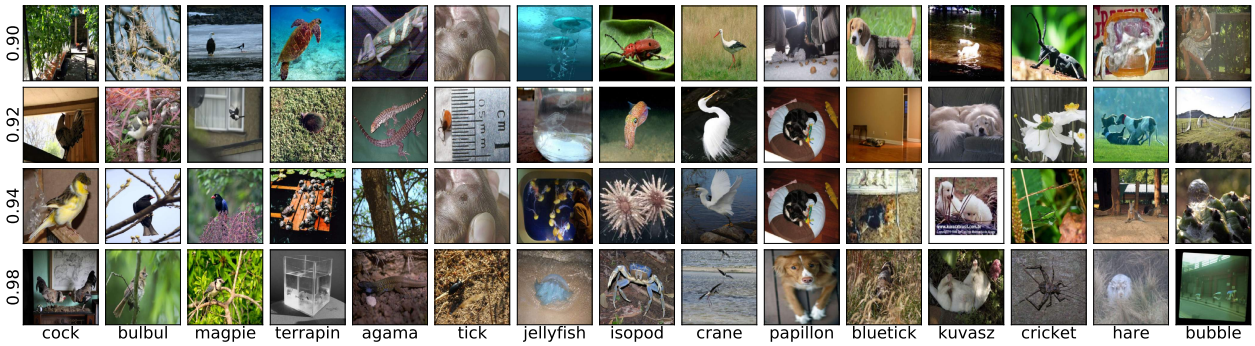
Figure 18. Examples of misclassified samples after masking the original well-trained model on CIFAR-10. The scores are estimated according to the uniform distribution so that weights are masked out randomly. (a) layer-wise masking scores (Ramanujan et al., 2020); (b) layer-wise masking weights directly; (c) model-wise masking scores; (d) model-wise masking weights directly. For layer-wise masking, a fixed mask ratio is set for every layer of the model, meaning that the same ratio of weights is masked out for every layer; for model-wise masking, the model is considered as a whole with all weights united first and then masked according to the mask ratio. When masking scores, we first generate a score for every weight according to the uniform distribution and then mask out those weights with smaller scores; when masking weights, we directly mask out weights with smaller magnitudes. The images show that masking weights can't efficiently detect atypical samples while masking scores can. For masking scores, masking with a smaller ratio forces the model to misclassify simple samples (clear contours around subjects, single color background) while masking with a larger ratio guide the model to misclassify complex samples (unclear contours, noisy background). This inspection empirically supports our claims that with proper mask ratio, we can detect atypical samples and therefore can force the model to forget them. Besides that, mask weights don't show any particular difference with different mask ratios, and thus it may be inappropriate for mining atypical samples.



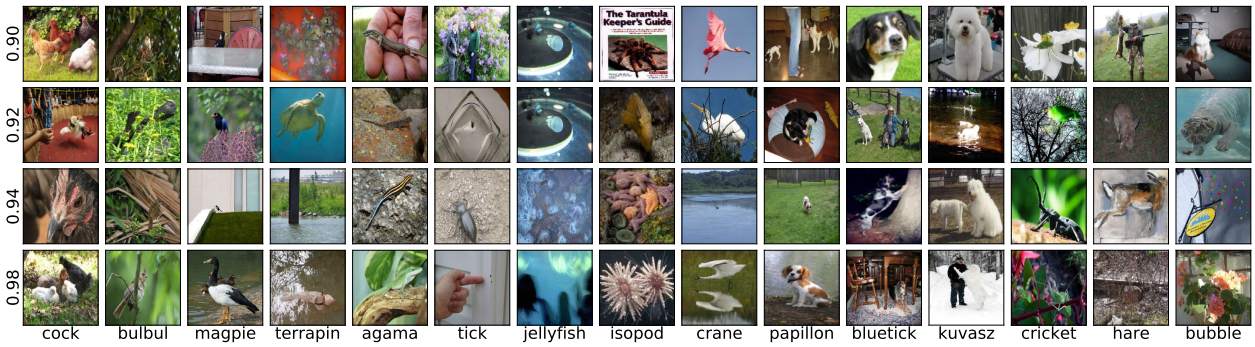
(a) Layer-wise masking scores



(b) Layer-wise masking weights



(c) Model-wise masking scores



(d) Model-wise masking weights

Figure 19. Examples of misclassified samples after masking the original well-trained model on ImageNet.