RESEARCH ARTICLE

AOE: Exhaustive Out-of-Distribution Detection via Recalibrating Outlier Labels

Fengqiang Wan[1], Qing-Yuan Jiang[1], Fu Shen[2], Yang Yang[1]✉

1. School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China.
2. Key Laboratory of Modern Agricultural Equipment, Ministry of Agriculture and Rural Affairs, P.R.China.

Received May 27, 2026; accepted July 3, 2026

E-mail: yyang@njust.edu.cn.

Abstract

Out-of-distribution (OOD) detection is essential for deploying machine learning models in open-world and safety-critical scenarios, where test inputs may deviate from the training distribution and overconfident predictions on unknown samples can lead to unreliable decisions. Outlier Exposure (OE) has emerged as a promising OOD detection paradigm by introducing auxiliary outliers during training to enlarge the margin between in-distribution (ID) and OOD samples. Existing OE-based methods typically enlarge this margin by employing uniform labels to maximize the entropy of OOD samples over ID categories. However, we theoretically show that uniform labels inevitably disregard the relations between OOD samples and ID categories, termed the over-softening effect, leading to a suboptimal margin bound. Our theoretical analysis further reveals that explicitly exploiting such relations can instead yield improved OOD detection performance. Motivated by this insight, we propose Adaptive Confidence OE (AOE), a simple yet effective method that leverages temperature scaling to recalibrate outlier labels. Specifically, AOE generates adaptive soft targets from temperature-scaled model predictions for OOD samples, where the learnable temperature smooths the prediction distribution without fully erasing class-wise relational information. By supervising OOD samples with these adaptive soft targets, AOE preserves the semantic proximity between OOD samples and ID categories while encouraging the softened targets to approach a high-entropy distribution, thereby suppressing overconfident OOD predictions and enlarging the separation margin. Extensive experiments across diverse benchmarks demonstrate the effectiveness of AOE. Compared with OE, AOE consistently lowers FPR95 by 2.40% and 2.51% on CIFAR-10, 0.64% and 9.11% on CIFAR-100, and 2.02% and 2.10% on ImageNet-200 under near-OOD and far-OOD settings, respectively.

Key words

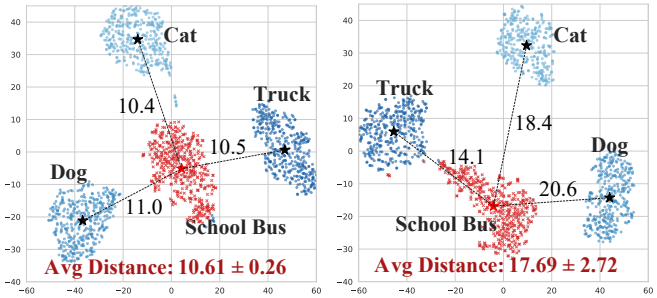
Out-of-distribution, outlier exposure, temperature scaling

1 Introduction

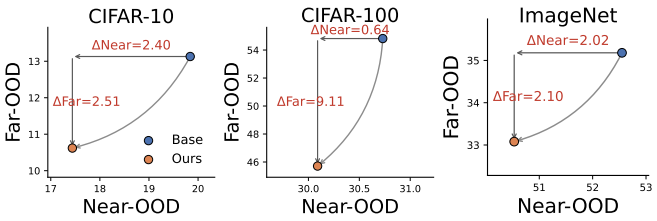
Machine learning models are typically predicated on the assumption that the test distribution coincides with that of the training distribution. However, this assumption often fails in real-world settings, where models are exposed to out-of-distribution (OOD) samples and suffer significant performance degradation, particularly in safety-critical applications [1–4]. To address this challenge, OOD detection [1, 5, 5–7] has emerged as an effective solution. It aims to accurately recognize in-distribution (ID) samples while identifying OOD samples [8–10].

Pioneering works like MSP [1] and Energy [7] are designed to develop score functions that can effectively distinguish OOD samples from ID data. Nevertheless, the fixed parameters of these models impose a significant constraint on their adaptability and overall performance. To tackle this issue, a variety of training-time regularization methods [11–13] such as LogitNorm [12] have been introduced. Training-time regularization methods aim to modify learning strategies so as to enhance the discriminative capacity of learned representations for improved OOD detection. Meanwhile, with the aid of additional auxiliary data, outlier exposure (OE)-based methods [5, 14, 15], e.g. OE [14] and DOE [5], facilitate the learning of more robust and generalizable decision boundaries. Among existing OOD detection methods, OE-based methods stand out as the most effective approach and consistently achieve superior performance as it leverages surrogate OOD samples during training to help the model distinguish between ID and OOD distributions.

Despite the effectiveness of OE-based methods, they inevitably disregard the relations of OOD samples with respect to different ID categories, which degrades OOD detection performance. We use a toy experiment to illustrate this issue. As shown in Fig. 1(a), where ID categories comprise *cat*, *dog*, and *truck*, while the OOD category is the *school bus*. To enlarge the margin between ID and OOD samples, OE-based methods assign uniform labels to OOD samples [14, 16], which inevitably gives rise to an *over-softening effect*, resulting in nearly uniform distances between the model predictions of OOD samples and those of different ID samples. In particular, OOD samples maintain a smaller distance to semantically irrelevant categories (e.g., *Cat*) than



(a). t-SNE of baseline (left) and ours (right).



(b). OOD detection performance (FPR95).

Fig. 1 (a) t-SNE visualization comparing the baseline (uniform labels) and our method (preserving relations) for the OOD category and ID categories. Compared with the baseline, our method exhibits a larger separation margin between ID and OOD samples. (b) The *over-softening effect* is consistently observed on CIFAR-10, CIFAR-100, and ImageNet, where our method leads to improved OOD detection performance.

to the most relevant one (e.g., *Truck*), as shown in Fig. 1(a) (left). Meanwhile, as shown in Theorem 1, disregarding such relations leads to a suboptimal separation bound, thereby degrading OOD detection performance. Intuitively, OOD samples often contain semantic information with certain ID categories [17–19]. A desirable model behavior is therefore to maintain relatively lower distance to the most semantically similar ID category compared with others [12]. Theorem 2 further shows that preserving relations mitigates the over-softening effect, as shown in Fig. 1(a) (right) and improves OOD detection, which is also observed empirically in Fig. 1(b).

Building on this insight, we propose Adaptive Confidence OE (AOE), a simple yet effective method that adaptively refines outlier labels via temperature scaling. Specifically, AOE employs a learnable temperature to smooth the model predictions for OOD samples and uses the resulting soft targets to supervise OOD training, thereby preserving the relations between OOD samples and ID categories; meanwhile, optimizing the temperature drives these smoothed targets toward a high-entropy distribution. To ensure robust convergence, AOE supports both joint optimization of model parameters and temperatures, as well as an alternating optimization strategy that decouples their respective update dynamics. Extensive experiments across diverse benchmarks demonstrate that AOE consistently outperforms existing state-of-the-art methods and can be seamlessly integrated with them.

The contributions of this work are summarized as follows:

- We provide a theoretical analysis showing that the commonly used uniform labeling strategy in OE induces an *over-softening effect*, which overlooks the intrinsic relationships between OOD samples and ID categories, leading to a suboptimal separation margin. We further demonstrate that introducing a temperature parameter T can effectively alleviate this issue.
- Motivated by this analysis, we propose AOE, a simple yet effective method that recalibrates outlier labels via temperature scaling. By leveraging temperature-scaled model predictions as soft targets, AOE preserves semantic relationships with ID classes while promoting high-entropy distributions, thereby improving margin separation.
- We conduct extensive experiments on multiple OOD detection benchmarks, demonstrating that AOE consistently improvements over corresponding OE-based methods and achieves robust gains across diverse datasets and model architectures.

2 Related Work

2.1 OOD Detection

OOD detection plays a vital role in open-world machine learning by accurately classifying ID samples while identifying OOD samples [1, 20]. Early post-hoc OOD detection methods [21,22] primarily focus on designing discriminative scoring functions to separate OOD samples at inference time. In contrast, training-time regularization methods [12, 23–26] aim to improve OOD detection by designing training strategies that encourage the learning of more discriminative representations.

Unlike the aforementioned methods, OE-based methods [14] incorporate realistic outliers into OOD detection models to enhance the robustness of decision boundaries. OE encourages OOD samples to exhibit a more uniform response during training, thereby reducing model overconfidence [14]. Building on this idea, MixOE [27] leverages MixUp between ID and OOD samples to synthesize more diverse OOD samples. DOE [5] further enhances robustness by generating hard OOD samples through model-driven transformations, enabling more reliable detection. DAL [16] synthesizes OOD samples by searching for the worst case within a Wasserstein ball around the OOD distribution, thereby reducing distribution discrepancy. OCL [28] explicitly assigns OOD samples to an additional $(K+1)$-th outlier class. By introducing outliers during training, these methods significantly improve OOD detection performance. However, these methods typically supervise OOD samples with uniform labels, which inevitably leads to the *over-softening effect*.

2.2 Temperature Scaling

Since the temperature coefficient controls the sharpness of the softmax output distribution, temperature scaling [29–31] has attracted significant attention. As a post-processing calibration technique, temperature scaling is first popularized for model calibration [29]. Temperature scaling [29] proposes an effective method to rescale the logits using a shared temperature parameter. T-CIL [32] optimizes the temperature by minimizing the calibration loss on adversarially perturbed samples, where perturbations are guided by feature-space distances. Additionally, temperature scaling has been widely applied in knowledge distillation [30] and contrastive learning [33]. Inspired by these practices, we incorporate temperature scaling to recalibrate outlier labels.

Table 1 Frequently used notations and their mathematical meanings.

Notation	Meaning
$\mathcal{X}$	Input space.
$\mathcal{Y} = \{y_1, \ldots, y_K\}$	Label set of K ID categories.
$\mathcal{D}_{\text{ID}}$, $\mathcal{D}_{\text{OOD}}$	ID and OOD distributions.
x_i, x_o	ID and OOD samples.
$f(\cdot;\theta)$	Classification model with parameters θ.
$s(\cdot)$	Softmax function.
$S(x;\theta)$	OOD confidence score.
$\Delta(\theta)$	Separation margin between ID and OOD samples.
U	Uniform distribution over ID categories.
$m(f(x;\theta))$	Prediction difference between the top two logits.
μ_i, μ_o, ν_o^2	Statistics of ID/OOD prediction differences.
T	Learnable temperature for outlier-label recalibration.
$q_T = s(f(x_o;\theta)/T)$	Temperature-scaled outlier target.

3 Methodology

In this section, we first introduce the preliminaries of OOD detection. We then provide a theoretical analysis showing that uniform outlier labels cause the over-softening effect and lead to a suboptimal separation margin. Motivated by this, we propose Adaptive Confidence OE to recalibrate outlier labels via temperature scaling, followed by the optimization strategy and its extension to different OE-based methods. The frequently used notations are summarized in Table 1.

3.1 Preliminary

Let $\mathcal{X}$ denote the input space and $\mathcal{Y} = \{y_1, \ldots, y_K\}$ the label set of K classes. The goal of OOD detection is to determine whether a sample $x \in \mathcal{X}$ originates from $\mathcal{D}_{\text{ID}}$ or $\mathcal{D}_{\text{OOD}}$. Given a model $f(\cdot;\theta)$ trained on ID and OOD samples, an OOD score function $S(x;\theta)$ assigns a confidence value to x. In general, the OOD detector $g_\lambda(x)$ is given by:

$$g_\lambda(x) = \begin{cases} \text{ID}, & \text{if } S(x;\theta) \geq \lambda, \\ \text{OOD}, & \text{otherwise.} \end{cases} \tag{1}$$

where λ is a threshold typically chosen to correctly classify the majority of ID samples (e.g., 95%). A sample is thus regarded as ID only if its confidence exceeds the threshold.

3.2 Theoretical Analysis in Outlier Exposure

OE-based methods improve OOD detection by introducing auxiliary outliers into the standard ID training. During training, these outliers are explicitly supervised with uniform labels to maximize entropy, thereby regularizing the decision boundary in regions beyond the support of the ID distribution [14]. Specifically, the training objective can be expressed as:

$$\min_{\theta} \; \mathcal{L}_{\text{ce}}(s(f(x_i;\theta))) \; + \; \alpha \mathcal{L}_{\text{KL}}\big(\mathcal{U} \,\big\|\, s(f(x_o;\theta))\big), \tag{2}$$

where $\mathcal{U} = (\frac{1}{K}, \ldots, \frac{1}{K})$ denotes the uniform distribution, and α is a hyperparameter balancing the contributions of ID samples x_i and OOD samples x_o. Here, $s(\cdot)$ represents the softmax function [1].

To analyze the impact of outliers on OOD detection performance, we introduce a definition to characterize the separability margin of ID samples and OOD samples.

Definition 1 (Separability Margin)**.** Separation between the $\mathcal{D}_{\text{ID}}$ and the $\mathcal{D}_{\text{OOD}}$ is quantified by the discrepancy between their induced confidence distributions. Specifically, the separation is defined as

$$\Delta(\theta) \triangleq \mathbb{E}_{x_i \sim \mathcal{D}_{\text{ID}}}[S(x_i;\theta)] - \mathbb{E}_{x_o \sim \mathcal{D}_{\text{OOD}}}[S(x_o;\theta)], \tag{3}$$

where $S(\cdot;\theta)$ denotes an OOD score from the model.

Based on Definition 1, effective OOD detection requires maintaining a sufficiently large separation margin between the confidence score distributions of ID and OOD samples [5,28]. However, the uniform supervision in Eq. (2) induces an inherent *over-softening effect* on model predictions for OOD samples. Since the relations between OOD samples and ID categories are reflected in the differences among their prediction scores [34, 35], we can characterize the effect at the logits level as follows.

Proposition 1. Let $z = f(x;\theta)$ denote the logits of x and z^u denote the logits with uniform supervision. The prediction differences across ID categories are contracted:

$$\left|z_i^u - z_j^u\right| \; \leq \; \left|z_i - z_j\right|, \forall\, i, j \in \{1, \ldots, K\},\; i \neq j. \tag{4}$$

Proposition 1 shows that uniform supervision systematically contracts prediction differences across ID categories, driving OOD predictions toward a uniform configuration.

Theorem 1. For any x, $k^\star(x) = \arg\max_k z_k$, and define the prediction differences $m(f(x;\theta)) \triangleq f_{k^\star(x)}(x;\theta) - \max_{j \neq k^\star(x)} f_j(x;\theta)$. Then the separation margin admits the following explicit upper bound:

$$\begin{aligned} \Delta(\theta) \leq \; & \frac{1}{1 + \exp\big(-\mu_i\big)} - \frac{1}{1 + (K-1)\exp\big(-\mu_o\big)} \\ & - \frac{(K-1)\exp\big(-\mu_o\big)}{2\big(1 + (K-1)\exp\big(-\mu_o\big)\big)^2}\, \nu_o^2 \; + \; \mathcal{R}. \end{aligned} \tag{5}$$

where $\exp(\cdot)$ denotes the exponential function, $\mathcal{R}$ collects higher-order remainder terms, and μ_i and μ_o denote the means of the corresponding prediction differences, respectively, with ν_o^2 representing the variance.

Theorem 1 reveals an explicit trade-off in the separation margin. Specifically, a moderate reduction of the prediction differences of OOD samples enlarges the separation margin via the first-order effect, whereas excessive contraction introduces a negative quadratic correction that diminishes the separation. This analysis highlights that effective OOD detection requires balanced regulation of prediction differences rather than their aggressive suppression.

Theorem 2. Consider the scaled labels $q_T = s(f(x_o;\theta)/T)$ with $T > 1$ for outlier x_o. Let z^t denote the logits after training with scaled supervision. Then the induced update of the differences satisfies

$$\begin{aligned} m(z^t) = \; & m(z) - \eta \Bigg[\frac{e^{m(z)} - 1}{e^{m(z)} + 1 + \sum_{k \neq a,b} \exp(z_k - z_b)} \\ & - \frac{e^{m(z)/T} - 1}{e^{m(z)/T} + 1 + \sum_{k \neq a,b} \exp\big((z_k - z_b)/T\big)} \Bigg]. \end{aligned} \tag{6}$$

where $a = \arg\max_k z_k$ and $b = \arg\max_{j \neq a} z_j$.

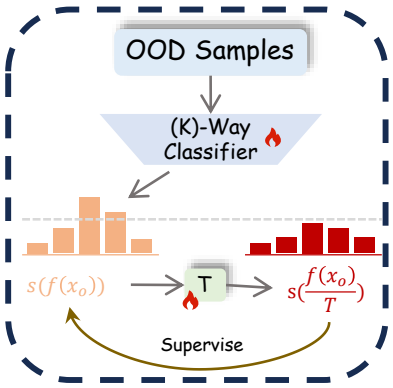


Fig. 2 Illustration of model training with AOE. During training, the model leverages a learned temperature to smooth its outputs on OOD samples and uses the resulting softened predictions as outlier labels. The model parameters and the temperature are optimized to achieve a self-evolving calibration of the decision boundary.

Corollary 1. Let $\Delta m_T \triangleq m(z) - m(z^t)$ denote the contraction. Under the conditions of Theorem 2, the contraction induced by temperature scaling at an optimal finite temperature $T^\star \in (1, \infty)$ is strictly smaller than that induced by uniform outlier exposure as $T \to \infty$.

Corollary 1 indicates that OE-based methods, corresponding to the extreme case $T \to \infty$, induces the strongest contraction. In contrast, a properly chosen finite temperature can smooth OOD predictions while preserving stronger relations. This observation motivates an adaptive mechanism for selecting the temperature to recalibrate outlier labels. Detailed proofs are provided in Appendix. Additionally, Appendix presents a theoretical analysis from the perspective of generalization bounds, substantiating the improved OOD detection capability.

3.3 Adaptive Confidence Outlier Exposure

Based on the above analysis, there exists an optimal temperature T^* that balances the trade-off between smoothing OOD predictions and preserving the prediction differences across ID categories. However, this optimal temperature cannot be computed explicitly during training. To this end, we treat the temperature coefficient as a learnable parameter within the OOD detection model. Fig. 2 illustrates the training framework of our proposal. During inference, the learned temperature T is not used to calibrate the model outputs.

Optimization Objective. Temperature T is optimized to regulate the model's prediction on OOD samples, encouraging the temperature-scaled predictions to align with a uniform distribution while aligning the model outputs with these softened targets so as to preserve the relation of OOD samples with respect to ID categories. Accordingly, the optimization objective is formulated as:

$$\begin{aligned}\min_{T} \ & \mathcal{L}_{\text{align}}\big(s(f(x_o;\theta)/T), \mathcal{U}\big) \\ & + \mathcal{L}_{\text{align}}\big(s(f(x_o;\theta)), s(f(x_o;\theta)/T)\big),\end{aligned} \tag{7}$$

where $\mathcal{L}_{\text{align}}(\cdot)$ denotes a distribution alignment loss, such as the Kullback–Leibler divergence [36]. In this formulation, the optimal temperature T^* is estimated from the current model predictions, while the learned temperature simultaneously influences the model outputs through temperature scaling [31,32]. Model parameters are optimized using available supervisory signals, where ID samples are supervised by ground-truth labels [37] and OOD samples are supervised by the smoothed predictions.

Algorithm 1 Training procedure of AOE

Require: ID dataset $\mathcal{D}_{\text{ID}}$, OOD dataset $\mathcal{D}_{\text{OOD}}$, model $f(\cdot;\theta)$, temperature T, trade-off parameter α
Ensure: Trained model parameter θ
1: **for** each training iteration **do**
2: Sample an ID mini-batch (x_i, y_i) from $\mathcal{D}_{\text{ID}}$
3: Sample an OOD mini-batch x_o from $\mathcal{D}_{\text{OOD}}$
4: Compute ID prediction $p_i = s(f(x_i;\theta))$
5: Compute OOD prediction $p_o = s(f(x_o;\theta))$
6: Compute temperature-scaled target $q_T = s(f(x_o;\theta)/T)$
7: Update θ and T by Eq. (8) or Eq. (9)
8: **end for**
9: **return** θ

Optimization Strategy. To determine the optimal model parameters θ and temperature T, we consider two optimization paradigms, namely *Joint Training* and *Alternating Training*.

Joint Training. This paradigm optimizes θ and T simultaneously within a single computational graph. Both variables are updated based on the combined objective:

$$\begin{aligned}\min_{\theta,T} \ & \mathcal{L}_{\text{ce}}\big(s(f(x_i;\theta)), y\big) + \alpha\Big(\mathcal{L}_{\text{align}}\big(s(f(x_o;\theta)/T), \mathcal{U}\big) \\ & + \mathcal{L}_{\text{align}}\big(s(f(x_o;\theta)), s(f(x_o;\theta)/T)\big)\Big)\end{aligned} \tag{8}$$

where $\mathcal{L}_{\text{ce}}(\cdot)$ denotes the standard cross-entropy loss [37], and α balances the contributions from ID and OOD samples [14].

Alternating Training. This paradigm decouples the optimization by iteratively updating T and θ. Specifically, T^* is first updated to satisfy the alignment constraint, and is subsequently treated as a constant to supervise θ:

$$\begin{aligned}\min_{\theta} \ & \mathcal{L}_{\text{ce}}\big(s(f(x_i;\theta)), y\big) + \alpha\ \mathcal{L}_{\text{align}}\big(s(f(x_o;\theta)), s(f(x_o;\theta)/T^*)\big) \\ \text{s.t.} \ & T^* \in \arg\min_{T} \mathcal{L}_{\text{align}}\big(s(f(x_o;\theta)/T), \mathcal{U}\big).\end{aligned} \tag{9}$$

In practice, Eq. (9) is solved by alternating minimization [38,39]. At iteration t, the temperature update is given by

$$T^{(t+1)} = T^{(t)} - \eta_T \nabla_T \mathcal{L}_{\text{align}}\big(s(f(x_o;\theta^{(t)})/T^{(t)}), \mathcal{U}\big), \tag{10}$$

where η_T denotes the learning rate for the temperature update. Given $T^{(t+1)}$, the model parameters are updated by

$$\begin{aligned}\theta^{(t+1)} = \theta^{(t)} - \eta_\theta \nabla_\theta \Big(& \mathcal{L}_{\text{ce}}\big(s(f(x_i;\theta^{(t)})), y_i\big) \\ & + \mathcal{L}_{\text{align}}\big(s(f(x_o;\theta^{(t)})), s(f(x_o;\theta^{(t)})/T^{(t+1)})\big)\Big),\end{aligned} \tag{11}$$

where η_θ denotes the learning rate for updating θ. In the parameter update step, $T^{(t+1)}$ is treated as a constant, yielding a first-order approximation to the optimization objective. The overall algorithm of our model is outlined in Algorithm 1.

Table 2 Comparison on ImageNet benchmarks. All values are percentages, and OOD detection results are averaged over multiple OOD test datasets. The best results are in **bold**, with the second-best underlined. Detailed results for each OOD dataset are provided in Appendix. ↑ indicates that larger values are better, while ↓ indicates that smaller values are better. *AOE-At* denotes alternating training, while *AOE-Jt* denotes joint training.

Method	Near-OOD		Far-OOD		Average		ID ACC↑
	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	
Post-hoc							
MSP	54.82±0.35	83.34±0.06	35.43±0.38	90.13±0.09	45.13±0.32	86.74±0.07	86.37±0.08
Energy	60.24±0.57	82.50±0.05	34.86±1.30	90.86±0.21	47.55±0.76	86.68±0.12	86.37±0.08
Training-time regularization							
LogitNorm	57.80±1.22	82.21±0.52	25.31±0.20	93.31±0.15	41.56±0.71	87.76±0.33	86.18±0.60
T2FNorm	55.65±0.20	82.72±0.05	25.25±0.48	93.38±0.11	40.45±0.33	88.06±0.07	86.52±0.21
UM	60.23±1.13	81.79±0.38	32.46±1.30	91.68±0.37	46.34±0.81	86.74±0.28	85.01±0.31
UMAP	60.81±0.84	81.08±0.39	32.47±0.67	91.62±0.29	46.64±0.74	86.35±0.32	86.37±0.08
PSKD	57.12±0.63	82.84±0.32	31.64±0.87	91.39±0.16	44.38±0.22	87.12±0.15	86.79±0.25
Outlier Exposure							
MixOE	86.79±0.25	82.57±0.23	40.12±0.66	88.39±0.02	49.03±0.44	85.49±0.12	85.73±0.09
DOE	54.14±0.51	83.23±0.60	37.60±2.95	88.24±1.45	45.87±1.72	85.73±1.00	79.87±3.12
DAL	51.85±0.40	85.17±0.05	35.76±0.14	88.55±0.05	43.81±0.27	86.86±0.05	86.18±0.29
OE	52.55±0.51	84.51±0.21	35.18±0.63	88.21±0.19	43.86±0.49	86.36±0.20	85.78±0.12
OCL	51.64±0.52	84.21±0.42	34.30±1.53	89.78±0.49	42.97±1.02	86.99±0.45	86.27±0.09
AOE-At	**50.53**±0.05	**85.61**±0.09	34.61±0.05	88.95±0.14	42.57±0.33	87.28±0.12	86.51±0.05
AOE-Jt	50.83±0.14	85.57±0.07	**33.08**±0.53	**89.24**±0.12	**41.95**±0.33	**87.41**±0.09	**86.49**±0.20

Table 3 OOD detection performance of multiple OE-based methods and their AOE-enhanced variants on CIFAR-10, CIFAR-100, and ImageNet-200. AOE generally improves the corresponding OE-based baselines in averaged near-OOD and far-OOD evaluations, with gains that are more pronounced in several far-OOD settings. These results demonstrate the effectiveness and general applicability of AOE as a plug-and-play enhancement for existing OE strategies.

ID Dataset	Type	Metric	OE	+AOE	MixOE	+AOE	DOE	+AOE	DAL	+AOE
CIFAR-10	Near-OOD	FPR95 ↓	19.84±0.95	**17.44**±0.05	51.45±7.78	**47.11**±5.05	20.39±0.15	**18.96**±0.78	20.91±0.71	**19.27**±0.79
		AUROC ↑	94.82±0.21	**95.64**±0.03	88.73±0.82	**89.07**±0.32	94.84±0.07	**95.17**±0.04	94.42±0.30	**94.69**±0.22
	Far-OOD	FPR95 ↓	13.13±0.53	**10.62**±1.50	33.84±4.77	**26.14**±1.80	15.59±1.47	**13.46**±1.21	21.40±2.56	**20.32**±0.27
		AUROC ↑	96.00±0.13	**97.04**±0.47	91.93±0.69	**93.61**±0.22	94.67±0.69	**96.16**±0.10	91.92±1.36	**92.12**±0.23
CIFAR-100	Near-OOD	FPR95 ↓	30.73±0.11	**30.09**±0.21	55.22±0.49	**54.70**±0.62	37.84±1.05	**32.21**±2.68	35.28±1.13	**32.16**±0.85
		AUROC ↑	88.30±0.10	**88.32**±0.10	80.95±0.20	**81.00**±0.31	86.61±0.29	**87.57**±1.29	85.97±0.07	**87.61**±0.33
	Far-OOD	FPR95 ↓	54.82±2.79	**46.12**±2.26	63.88±2.48	**60.33**±3.16	45.38±0.52	**43.64**±3.87	47.33±1.28	**41.59**±1.59
		AUROC ↑	81.41±1.49	**86.04**±0.38	76.40±1.44	**78.16**±1.05	84.30±0.81	**86.56**±3.88	82.42±0.26	**85.80**±0.78
ImageNet-200	Near-OOD	FPR95 ↓	52.55±0.51	**50.83**±0.14	57.95±0.23	**54.82**±3.47	54.14±0.51	**50.34**±0.11	51.85±0.40	**50.78**±0.36
		AUROC ↑	84.51±0.21	**85.57**±0.07	82.57±0.23	**83.83**±1.34	83.23±0.60	**85.60**±0.01	85.17±0.05	**86.20**±0.06
	Far-OOD	FPR95 ↓	35.18±0.63	**33.08**±0.53	40.12±0.66	**37.98**±3.23	37.60±2.95	**34.68**±0.02	35.76±0.14	**34.60**±0.10
		AUROC ↑	88.21±0.19	**89.24**±0.12	88.39±0.02	**88.67**±0.18	88.24±1.45	**88.81**±0.25	88.55±0.05	**88.60**±0.06

Extension to Different OE Strategies. AOE seamlessly adapts to diverse OE strategies. For methods focused on generating synthetic OOD samples [16], we directly adopt the objective defined in Eq. (7). Conversely, when leveraging auxiliary outlier classes [28], we align the predictive distribution over ID categories with a softened target distribution obtained via temperature scaling. Accordingly, Eq. (7) can be reformulated as follows:

$$\begin{aligned}\min_{T}\ & \mathcal{L}_{\text{align}}\big(s(f_{1:K}(x_o;\theta)/T), \mathcal{U}\big) \\ & + \mathcal{L}_{\text{align}}\big(s(f_{1:K}(x_o;\theta)), s(f_{1:K}(x_o;\theta)/T)\big),\end{aligned} \tag{12}$$

where $f_{1:K}(\cdot)$ denotes the logits corresponding to the ID categories and K represents the number of ID categories.

4 Experiments

4.1 Experimental Setup

To ensure a fair comparison, we follow established protocols and assess our approach on the OpenOODv1.5 benchmark [40]. The evaluation spans both small-scale and large-scale datasets, encompassing near-OOD characterized by semantic shifts as well as far-OOD involving more pronounced covariate shifts.

Table 4 Comparison on CIFAR benchmarks. All values are percentages, and OOD detection results are averaged over multiple OOD test datasets. The best results are in **bold**, with the second-best underlined. Detailed results for each OOD dataset are provided in Appendix. ↑ indicates that larger values are better, while ↓ indicates that smaller values are better. *AOE-At* denotes alternating training, while *AOE-Jt* denotes joint training.

Method	CIFAR-10					CIFAR-100				
	Near-OOD		Far-OOD		ID ACC↑	Near-OOD		Far-OOD		ID ACC↑
	FPR95↓	AUROC↑	FPR95↓	AUROC↑		FPR95↓	AUROC↑	FPR95↓	AUROC↑	
Post-hoc										
MSP	48.17±3.92	88.03±0.25	31.72±1.84	90.73±0.43	**95.06±0.30**	54.80±0.33	80.27±0.11	58.70±1.06	77.76±0.44	**77.25±0.10**
Energy	61.34±4.63	87.58±0.46	41.69±5.32	91.21±0.92	95.06±0.30	55.62±0.61	80.91±0.08	56.59±1.38	79.77±0.61	77.25±0.10
Training-time regularization										
LogitNorm	29.34±0.81	92.33±0.08	13.81±0.20	96.74±0.06	94.30±0.25	62.89±0.57	78.47±0.31	53.61±3.45	81.53±1.26	76.34±0.17
T2FNorm	26.47±0.35	92.79±0.13	12.75±0.73	96.98±0.23	94.69±0.07	58.47±1.35	79.84±0.40	51.25±2.52	82.73±1.01	76.43±0.13
UM	33.12±0.47	90.60±0.37	22.95±1.65	93.67±0.68	92.33±0.41	65.86±2.06	77.14±0.62	51.90±2.04	81.63±1.70	72.21±0.43
UMAP	33.01±0.06	91.00±0.07	21.70±1.57	94.20±0.36	95.06±0.30	59.71±0.65	79.49±0.23	52.11±2.36	81.62±1.37	77.25±0.10
PSKD	31.67±0.78	91.71±0.16	20.48±1.30	94.56±0.29	95.14±0.08	54.83±2.79	81.45±0.44	51.56±2.39	82.40±0.52	77.44±0.09
Outlier Exposure										
MixOE	51.45±7.78	88.73±0.82	33.84±4.77	91.93±0.69	94.55±0.32	55.22±0.49	80.95±0.20	63.88±2.48	76.40±1.44	75.13±0.06
DOE	20.39±0.15	94.84±0.07	15.59±1.47	94.67±0.69	94.32±0.19	37.84±1.05	86.61±0.29	**45.38±0.52**	84.30±0.81	75.69±0.26
DAL	20.91±0.71	94.42±0.30	21.40±2.56	91.92±1.36	92.95±0.54	35.28±1.13	85.97±0.07	47.33±1.28	82.42±0.26	75.60±0.34
OE	19.84±0.95	94.82±0.21	13.13±0.53	96.00±0.13	94.63±0.26	30.73±0.11	88.30±0.10	54.82±2.79	81.41±1.49	76.84±0.42
OCL	22.64±0.88	94.84±0.24	14.44±1.21	96.71±0.30	93.83±0.32	30.49±0.47	**88.91±0.16**	51.10±3.12	83.67±1.26	75.96±0.20
AOE-At	18.37±0.57	95.21±0.09	11.88±0.15	96.25±0.42	94.82±0.19	30.56±0.14	88.24±0.07	45.71±2.88	85.97±1.06	76.21±0.01
AOE-Jt	**17.44±0.05**	**95.64±0.03**	**10.62±1.50**	**97.04±0.47**	94.79±0.15	**30.09±0.21**	88.32±0.10	46.12±2.26	**86.04±0.38**	76.90±0.14

Datasets: Small-scale evaluations are conducted using CIFAR-10 and CIFAR-100 [41] as in-distribution datasets. The near-OOD datasets include CIFAR-100/10 and Tiny ImageNet [42], while the far-OOD datasets comprise MNIST [43], SVHN [44], Textures [45], and Places365 [46]. In addition, Tiny ImageNet-597 [42], with no category overlap with CIFAR-10/100 or the OOD datasets, is employed as a realistic outlier set following the OpenOOD evaluation protocol. For large-scale experiments, ImageNet-200, consisting of 200 categories sampled from ImageNet-1K [47], serves as the in-distribution dataset. The near-OOD evaluation utilizes SSB-hard [48] and NINCO [49], whereas the far-OOD evaluation considers iNaturalist [50], Textures [45], and OpenImage-O [51]. Furthermore, the remaining 800 classes from ImageNet-1K are used as realistic outliers.

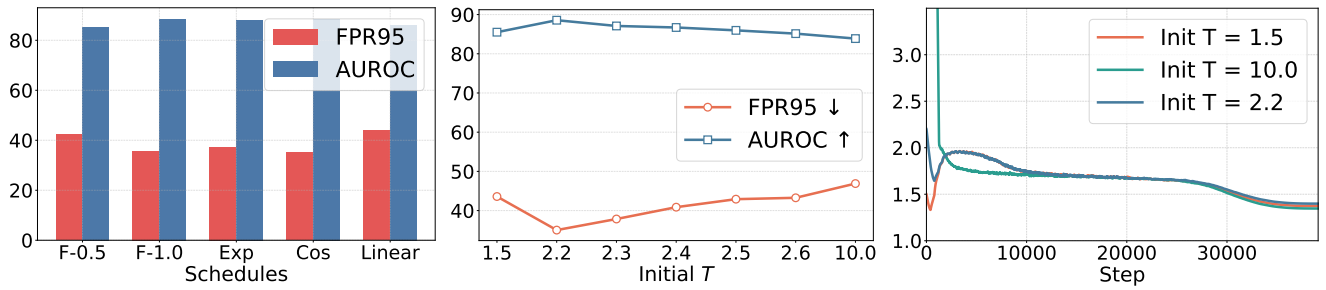
Implementation Details: To ensure fair evaluation, we follow the standardized OOD detection protocol provided by the OpenOOD benchmark [40]. ResNet-18 [52] is adopted as the primary backbone. All models are trained for 100 epochs using stochastic gradient descent with an initial learning rate of 0.1, cosine annealing scheduling [53], a momentum coefficient of 0.9, and a weight decay of 5×10^{-4}. The batch size is set to 128 for CIFAR-10 and -100, and 256 for ImageNet-200. During training, the temperature T is constrained within the range $[1.0, 10]$, and values outside this interval are clipped accordingly.

Evaluation Metrics: Following the setting of [14], we employ FPR95, AUROC and ID ACC for evaluation, where FPR95 refers to the false positive rate of OOD samples when the true positive rate of ID samples is fixed at 95%, and AUROC is a widely used metric that quantifies the overall separability between ID and OOD samples, based on the receiver operating characteristic curve.

Baselines: We select three categories of OOD detection methods for experiments: (1) Post-hoc methods, including MSP [1] and Energy [7]; (2) Training-time regularization methods, including LogitNorm [12], T2FNorm [23], UM/UMAP [25], and PSKD [26]; (3) Outlier exposure methods, including OE [14], MixOE [27], DAL [16], DOE [5] and OCL [28]. Furthermore, we explore the impact of different score functions, including ODIN [54], ASH [8], REACT [55], DICE [9], SCALE [56], SHE [57], GEN [58] and RankFeat [59].

4.2 Main Results

We report the OOD detection performance of AOE and SOTA methods on large-scale datasets in Table 2 and small-scale datasets in Table 4. For all methods, the mean performance and standard deviation are computed over multiple independent runs with different random initializations. Several observations can be drawn from the experimental results. (1) OE-based methods, including our proposed AOE, generally outperform training-time regularization methods and score-based methods. This advantage can be primarily attributed to the incorporation of outlier samples during training, which enhances the model's ability to discriminate OOD samples from ID samples. (2) However, conventional OE-based methods often compromise ID classification performance and fail to fully exploit the relations between OOD samples and ID categories, as theoretically analyzed in Theorem 1, particularly when assigning an additional class to OOD samples. In contrast, as stated in Theorem 2, our method effectively mitigates this over-

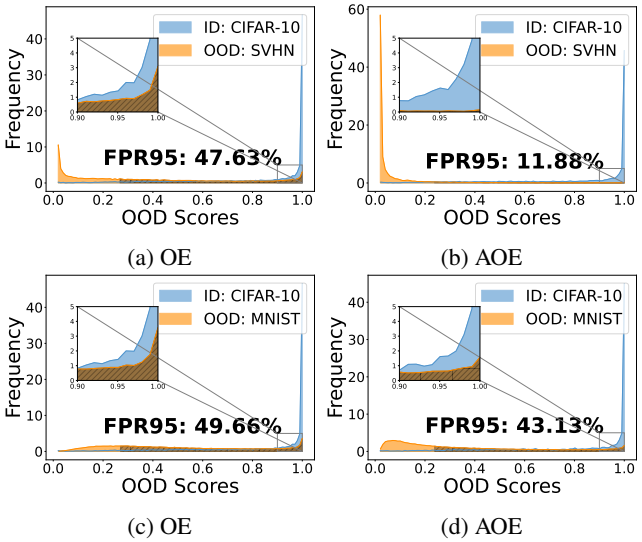


(a) Adjustment strategies of α (b) Temperature initialization (c) Evolution of T

Fig. 3 Ablation study on key design choices of AOE. CIFAR-100 is used as the ID dataset. (a) Effect of different adjustment strategies for α. (b) Impact of temperature initialization. (c) Evolution of the learned temperature T during training, showing a consistent convergence pattern.

Table 5 OOD detection performance with a series of fixed temperature values on CIFAR-10 and CIFAR-100 using MSP. The best results are highlighted in **bold**, and the second-best are underlined. ↑ indicates higher is better, while ↓ indicates lower is better.

Method	CIFAR-10					CIFAR-100				
	Near-OOD		Far-OOD		ID ACC↑	Near-OOD		Far-OOD		ID ACC↑
	FPR95↓	AUROC↑	FPR95↓	AUROC↑		FPR95↓	AUROC↑	FPR95↓	AUROC↑	
Random-Hard	20.54±0.13	94.76±0.20	15.47±0.91	95.44±0.69	94.53±0.19	30.99±0.36	**88.49±0.13**	51.23±2.16	83.60±1.29	76.79±0.40
Random-Soft	19.20±1.09	95.01±0.35	14.04±0.61	95.38±0.27	94.48±0.25	30.39±0.19	88.36±0.11	50.19±4.28	84.11±2.23	**77.10±0.31**
Fix T = 3.5	19.48±0.84	94.94±0.30	14.01±1.08	95.98±0.33	94.53±0.08	30.71±0.51	88.26±0.19	51.57±1.79	83.60±0.96	76.81±0.38
Fix T = 4.5	18.70±0.93	95.07±0.31	11.35±1.24	96.61±0.42	94.50±0.26	31.92±1.82	87.49±0.99	46.30±9.58	85.48±3.89	76.35±0.98
Fix T = 5.5	19.37±1.08	94.93±0.35	13.89±0.65	95.63±0.51	94.66±0.11	32.42±2.31	87.41±1.01	**45.64±12.77**	84.99±5.14	76.39±0.97
AOE-At	18.37±0.57	95.21±0.09	11.88±0.15	96.25±0.42	**94.82±0.19**	30.56±0.14	88.24±0.07	45.71±2.88	85.97±1.06	76.21±0.01
AOE-Jt	**17.44±0.05**	**95.64±0.03**	**10.62±1.50**	**97.04±0.47**	94.79±0.15	**30.09±0.21**	88.32±0.10	46.12±2.26	**86.04±0.38**	76.90±0.14



(a) OE (b) AOE

(c) OE (d) AOE

Fig. 4 OOD score distributions of OE and AOE. CIFAR-10 is used as the ID dataset, while SVHN and MNIST are used as OOD datasets. The histograms compare the confidence score distributions of OOD samples under OE and AOE, with emphasis on the high-score region. Compared with OE, AOE produces fewer high-confidence OOD predictions and achieves lower FPR95, indicating improved separation between ID and OOD samples.

softening issue, leading to improved overall performance. As a result, compared with existing methods, AOE achieves superior performance across nearly all evaluated scenarios. (3) *AOE-Jt* of the temperature parameter T outperforms *AOE-At*, as it allows the model to co-adapt the temperature alongside its feature representations, achieving more accurate alignment with ID decision boundaries. Additional plug-in results on existing OE methods are shown in Table 3, confirming the generality and compatibility of AOE.

4.3 Ablation Study

Impact of Adjustment Strategies for α. Balancing coefficient α serves as the most influential factor for our model's effectiveness. As discussed in Section 3.3, we investigate strategies for adjusting the value of α during different stages of training. We designed four scheduling strategies for the parameter α: (1) a fixed-value strategy, i.e., $\alpha_t = c$, where t and c denote the epoch and constant; (2) an exponential scheduling strategy, i.e., $\alpha_t = 1 - e^{-t/35}$; (3) a cosine annealing strategy, i.e., $\alpha_t = 0.5 - \cos((t+1)\pi/100)/2$; and (4) a linear scheduling strategy, i.e., $\alpha_t = \min\{1, (t+1)/100\}$. For the first strategy, we set $c = \{0.5, 1\}$ for experiments. The results are shown in Fig. 3(a), where "F-0.5" and "F-1.0" denote the fixed-value strategy with different c, "Exp", "Cos", and "Linear" respectively denote the exponential scheduling strategy, cosine annealing strategy and linear scheduling strategy. We can find that different scheduling strategies for α result in comparable performance, with the cosine annealing strategy achieving the best overall results.

Impact of Temperature Initialization. Temperature T is treated as a learnable parameter. We report the OOD detection performance with different initialization values for $T \in [1.5, 10]$ in Fig. 3(b). Furthermore, we illustrate the change of T during model training in Fig. 3(c).

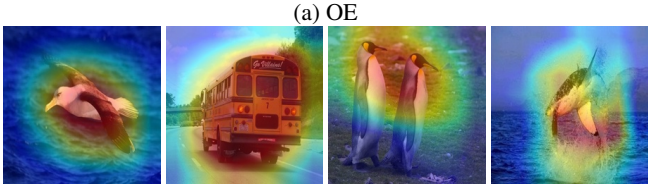
Table 6 OOD detection performance of OE and its AOE-enhanced variant across different backbone architectures on CIFAR-100. AOE consistently improves performance across architectures, with more significant gains observed in far-OOD settings

Backbone	Near-OOD				Far-OOD			
	FPR95 ↓		AUROC ↑		FPR95 ↓		AUROC ↑	
	OE	AOE	OE	AOE	OE	AOE	OE	AOE
WideResNet	30.27±0.81	30.02±0.36	89.38±0.15	89.34±0.06	36.30± 0.58	35.04±0.82	88.92± 0.36	89.68±0.32
ResNet-18	30.73±0.11	30.09±0.21	88.30±0.10	88.32±0.10	54.82±2.79	46.12±2.26	81.41±1.49	86.04±0.38
DenseNet	32.04±0.42	31.62±0.52	87.05±0.34	86.81±0.02	52.34±2.37	50.21±0.32	81.84±0.57	82.73±1.04

Table 7 OOD detection performance across different scoring functions on CIFAR-100. AOE consistently improves OOD detection across a wide range of scoring methods, achieving lower FPR95 and higher AUROC in most cases, with particularly significant gains in far-OOD scenarios.

OOD Scores	Near-OOD						Far-OOD					
	FPR95 ↓			AUROC ↑			FPR95 ↓			AUROC ↑		
	OE	*AOE-At*	*AOE-Jt*	OE	*AOE-At*	*AOE-Jt*	OE	*AOE-At*	*AOE-Jt*	OE	*AOE-At*	*AOE-Jt*
Energy	30.34±0.21	30.49±0.33	30.28±0.71	88.42±0.05	88.25±0.21	88.18±0.37	50.59±0.76	39.26±2.26	41.10±1.38	84.10±0.85	88.34±0.86	87.86±0.28
ODIN	40.22±0.66	39.09±2.58	38.46±0.23	86.19±0.51	86.46±0.78	86.82±0.13	49.08±3.40	45.79±2.25	48.96±1.75	84.78±1.32	86.70±0.38	85.93±0.45
ASH	36.46±1.41	35.83±0.79	36.88±0.84	85.74±1.23	85.58±0.62	85.08±0.66	55.53±0.84	42.49±2.14	44.72±1.76	82.78±1.01	87.10±0.92	86.29±0.62
REACT	31.09±1.39	33.37±2.10	31.46±1.64	87.92±0.49	86.80±1.33	87.60±0.62	49.62±0.98	40.33±0.59	40.72±2.51	84.36±0.85	88.19±0.49	88.03±0.37
DICE	31.44±0.12	31.32±0.66	31.08±1.30	87.32±0.16	87.34±0.32	87.21±0.50	49.45±0.63	37.10±1.93	40.14±0.83	84.25±0.64	88.38±0.88	87.63±0.15
SCALE	36.98±0.29	30.74±0.94	32.86±3.30	85.71±0.10	88.11±0.32	87.13±1.17	55.49±1.52	38.27±1.89	41.16±3.64	83.43±0.07	88.69±0.84	87.85±0.94
SHE	32.47±0.26	31.94±1.05	32.31±1.43	86.67±0.15	86.81±0.46	86.35±0.65	55.92±1.44	41.86±1.64	44.47±1.30	82.24±1.02	87.13±0.82	86.44±0.13
GEN	30.18±0.31	30.48±0.32	30.25±0.72	88.54±0.21	88.37±0.23	88.20±0.37	51.05±1.11	40.01±2.98	41.14±1.40	83.97±0.80	88.24±0.98	87.86±0.29
RankFeat	50.38±10.08	48.37±1.81	50.33±0.68	77.64±4.37	76.86±1.16	77.78±1.83	90.43±3.29	65.95±12.66	70.65±13.77	55.36±8.49	73.44±8.62	68.82±12.74

(a) OE



(b) AOE

Fig. 5 Grad-CAM visualizations for OOD samples under OE and AOE. Compared with OE, AOE produces more focused and semantically meaningful activation regions, indicating improved uncertainty modeling for OOD samples.

The results reveal that the initial value of T significantly affects the final performance. Interestingly, regardless of the initialization, the learned temperature consistently follows a similar trajectory, converging to a range between 1.5 and 2, and then gradually decreasing towards the end of training. This indicates that early-stage temperature values play a critical role in guiding effective optimization.

4.4 Further Analysis

Impact of Learnable Temperature. We evaluate the impact of a learnable temperature by comparing AOE with alternative pseudo-labeling strategies, including random hard labels, random soft labels, and fixed temperature coefficients. The results indicate that AOE's performance gains do not stem from arbitrary label smoothing, but from preserving relations during calibration. As shown in Table 5, random hard and soft labels yield only marginal improvements over standard OE, while fixed-temperature calibration offers moderate gains but suffers from limited robustness. In contrast, AOE, with adaptive pseudo-labels and a learnable temperature, consistently outperforms all baselines, underscoring the effectiveness of relations-preserving calibration for reliable OOD detection.

Impact of OOD Score Function. Since our study is orthogonal to the choice of score functions, we investigate the impact of different scoring functions on the final OOD detection performance in this section. We report the results on the CIFAR-100 benchmark with various score functions in Table 7. We observe that standard OE-based methods exhibit degraded performance when combined with different scoring functions. This degradation arises because these methods distort the model's confidence landscape, resulting in misaligned score distributions at test time. In contrast, AOE achieves consistently strong performance across diverse scoring functions by preserving relations during training, thereby ensuring better compatibility with OOD score.

Impact of Auxiliary Outlier Sources. To examine the influence of auxiliary outlier sources, we evaluate two representative sources widely used in OE-based studies, including 300K random images sampled from 80 Million Tiny Images [60] and ImageNet-RC [61]. As shown in Fig. 7(a), AOE maintains a consistent performance advantage over OE under different auxiliary outlier sources. This observation suggests that the improvement is not merely caused by a particular outlier exposure dataset, but is more closely related to the proposed recalibration of outlier labels, which adaptively preserves the predictive relations between auxiliary outliers and ID categories.

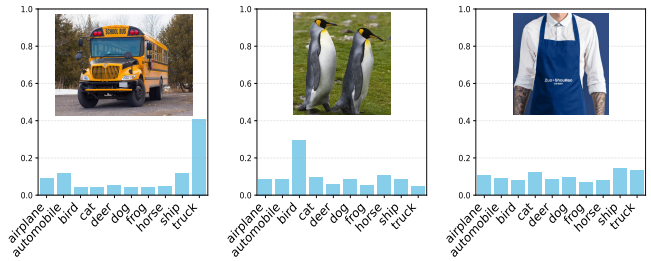
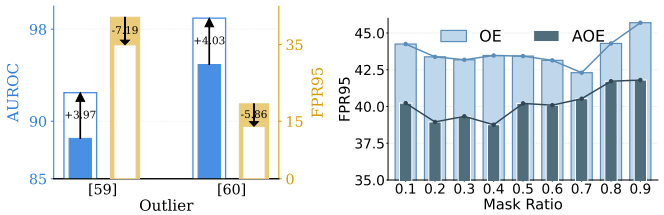


Fig. 6 Case studies of temperature-scaled outlier targets generated by AOE. Each bar plot shows the soft target distribution over CIFAR-10 ID categories for a representative OOD sample. AOE assigns relatively higher probabilities to semantically related ID categories while maintaining uncertainty for ambiguous OOD samples.



(a) Auxiliary outlier sources (b) Number of auxiliary outliers

Fig. 7 Further analyses of auxiliary outliers on CIFAR-100. (a) Effect of auxiliary outlier sources. The upward arrows indicate the AUROC increasing from OE to AOE, whereas the downward arrows denote the reduction in FPR95. (b) Effect of the number of auxiliary outliers. AOE consistently outperforms OE in both settings.

Impact of the Number of Auxiliary Outliers. We further investigate the robustness of AOE to the number of exposed auxiliary outliers. Specifically, we randomly discard auxiliary OOD samples according to different mask ratios and compare AOE with OE under the same training protocol. As shown in Fig. 7(b), AOE consistently outperforms OE across different numbers of auxiliary outliers. Moreover, the results indicate that simply increasing the number of exposed outliers does not necessarily lead to the best performance, suggesting that the informativeness of auxiliary outliers is also crucial.

Impact of Model Architecture. As shown in Table 6, AOE consistently improves over OE across all backbones, with more pronounced gains on far-OOD. The improvements are especially significant for ResNet-18, while remaining stable for stronger architectures such as WideResNet and DenseNet. These results indicate that AOE is robust to architectural choices and particularly beneficial for weaker models and harder OOD settings.

Visualization. We further present the histograms of OOD scores for our proposed method and the OE baseline for comparison. Fig. 4(a) and Fig. 4(b) use CIFAR-10 as the ID dataset and SVHN as the OOD dataset, whereas Fig. 4(c) and Fig. 4(d) use MNIST as the OOD dataset. OE assigns uniform targets to OOD samples, enforcing equal class uncertainty and resulting in suppressed the model's ability to distinguish ambiguous samples and results in overconfident responses. Fig. 4 shows substantial overlap between ID and OOD scores, indicating poor separability. In contrast, AOE replaces uniform supervision with adaptive pseudo-labels modulated by a learnable temperature, which introduces sample-dependent uncertainty into the training signal. This mechanism encourages calibrated confidence for hard OOD samples, leading to a noticeable reduction in high-confidence OOD predictions and a clearer separation between ID and OOD score distributions.

Case Study. We examine the pseudo-label distributions of several real OOD samples with respect to ID categories to gain deeper insight into the behavior of our method. The results are presented in Fig. 6. For the OOD sample labeled as "School Bus," our method assigns relatively high confidence to the semantically related ID class "Truck." A similar pattern is observed for the sample labeled as "King Penguin." In contrast, for samples such as "Apron", the model produces a more uniform distribution over ID categories, reflecting higher uncertainty. We further visualize the corresponding Grad-CAM [62] activation maps in Fig. 5 to examine how different supervision strategies affect the model's attention. Under OE, the attention is often diffuse, reflecting the absence of meaningful guidance from uniform target supervision. In contrast, AOE produces more focused and semantically coherent attention regions, aligning well with the structure of the learned pseudo-labels.

Computational Overhead. To further evaluate the practical efficiency of AOE, we analyze the additional overhead introduced by the proposed objective. As defined in Eq. (7), AOE only introduces a learnable temperature T to recalibrate outlier labels from the original OOD logits. This design keeps the model architecture unchanged and reuses the same backbone computation as standard OE. The extra training cost mainly comes from temperature scaling and distribution alignment on existing OOD logits, which is minor compared with the standard forward and backward propagation. Moreover, since T is used only for training-time label recalibration and is discarded during inference, AOE has the same inference procedure and inference cost as the corresponding OE-based method.

■ 5 Conclusion

We identify a fundamental limitation of conventional Outlier Exposure, where uniform supervision on OOD samples induces excessive margin contraction and restricts distributional separation. To address this issue, we propose AOE, which replaces uniform targets with adaptive pseudo-labels derived from temperature-scaled model predictions. Our theoretical analysis establishes a direct connection between the relations of OOD samples with respect to ID categories and OOD separability, and shows that adaptive temperature scaling effectively mitigates the over-softening effect. We further introduce joint and alternating optimization schemes to learn the temperature during training. Extensive experiments demonstrate that AOE consistently improves OOD detection performance while preserving competitive ID Acc, and can be seamlessly integrated with existing OE strategies.

Limitations and Future Work. Despite these improvements, the effectiveness of AOE is closely related to the presence of meaningful predictive relations between auxiliary outliers and ID categories. When auxiliary outliers are substantially distant from the ID label space, the resulting temperature-scaled pseudo-labels may approach uniform targets, thereby reducing the distinguishable advantage over conventional OE. This indicates that AOE is particularly effective when

auxiliary outliers contain partial semantic proximity with ID classes. Future work may further strengthen AOE through relation-aware outlier weighting, enabling informative outliers to be more effectively exploited while weakly related samples are adaptively suppressed.

■ **Acknowledgements**

This work was supported by the NSFC(62276131, 62506168), the Natural Science Foundation of Jiangsu Province of China (BK20240081, BK20251431), the Fundamental Research Funds for the Central Universities (No. 30925010205), the Special Research Project on Teaching Reform of General Artificial Intelligence Courses in Jiangsu Undergraduate Universities (No. ZNT-10), and the Open Project of the Key Laboratory of Modern Agricultural Equipment, Ministry of Agriculture and Rural Affairs, P.R. China.

■ **Competing interests**

The authors declare that they have no competing interests or financial conflicts to disclose.

■ **Appendixes**

Theoretical proof

Proof of Proposition 1

We analyze the parameter update under the Outlier Exposure objective using a uniform target $\mathcal{U} = (1/K, \ldots, 1/K)$. Let $z = f(x;\theta) \in \mathbb{R}^K$ be the logits for a given sample x, and $p = s(z)$ be the predicted softmax probabilities, where $p_k = \frac{e^{z_k}}{\sum_{j=1}^K e^{z_j}}$. The KL divergence loss with respect to the uniform distribution is defined as:

$$\mathcal{L}_{\text{KL}} = -\frac{1}{K}\sum_{k=1}^{K} \log p_k + \text{const.} \tag{13}$$

The gradient of the loss with respect to the k-th logit is:

$$\frac{\partial \mathcal{L}_{\text{KL}}}{\partial z_k} = p_k - \frac{1}{K}. \tag{14}$$

Under a gradient descent update with learning rate η, the updated logits are $z^+ = z - \eta \nabla_z \mathcal{L}_{\text{KL}}$. For any two categories i and j, the difference between their updated logits is:

$$z_i^+ - z_j^+ = (z_i - z_j) - \eta\left(\frac{\partial \mathcal{L}_{\text{KL}}}{\partial z_i} - \frac{\partial \mathcal{L}_{\text{KL}}}{\partial z_j}\right) \tag{15}$$

$$= (z_i - z_j) - \eta(p_i - p_j). \tag{16}$$

Assume $z_i > z_j$ without loss of generality. Due to the strict monotonicity of the softmax function, $z_i > z_j \implies p_i > p_j$. Thus, $p_i - p_j > 0$. For a sufficiently small η such that $z_i^+ - z_j^+$ maintains the same sign as $z_i - z_j$, we have:

$$0 < z_i^+ - z_j^+ < z_i - z_j. \tag{17}$$

Taking the absolute value, we obtain:

$$|z_i^+ - z_j^+| = |(z_i - z_j) - \eta(p_i - p_j)| \le |z_i - z_j|. \tag{18}$$

This demonstrates that optimizing toward a uniform distribution inherently contracts the relative differences between logits, resulting in the over-softening effect.

Proof of Theorem 1

Let $z = f(x;\theta) \in \mathbb{R}^K$ denote the logits of an input x, and define $k^\star(x) = \arg\max_{k\in\{1,\ldots,K\}} z_k$ and $m(x;\theta) \triangleq z_{k^\star(x)} - \max_{j\neq k^\star(x)} z_j$. We adopt the widely used maximum softmax probability (MSP) as the OOD score:

$$S(x;\theta) \triangleq \max_k \text{softmax}(z)_k = \frac{\exp(z_{k^\star(x)})}{\sum_{j=1}^K \exp(z_j)}. \tag{19}$$

For convenience, define the ID and OOD random variables $m_i \triangleq m(x_i;\theta), x_i \sim \mathcal{D}_{\text{ID}}$ and $m_o \triangleq m(x_o;\theta), x_o \sim \mathcal{D}_{\text{OOD}}$, with corresponding statistics $\mu_i \triangleq \mathbb{E}[m_i]$, $\mu_o \triangleq \mathbb{E}[m_o]$, and $v_o^2 \triangleq \text{Var}(m_o)$.

The distribution separation margin is defined as

$$\Delta(\theta) = \mathbb{E}_{x_i\sim\mathcal{D}_{\text{ID}}}[S(x_i;\theta)] - \mathbb{E}_{x_o\sim\mathcal{D}_{\text{OOD}}}[S(x_o;\theta)]. \tag{20}$$

We first establish a tight connection between the MSP score and the margin $m(x;\theta)$.

Lemma 1 (Margin bounds for MSP)**.** For any x with margin $m(x;\theta)$, the MSP score satisfies

$$\frac{1}{1+(K-1)\exp(-m(x;\theta))} \le S(x;\theta) \le \frac{1}{1+\exp(-m(x;\theta))}. \tag{21}$$

Proof. Let $\delta_j = z_{k^\star} - z_j$ for $j \neq k^\star$, so that

$$S(x;\theta) = \frac{1}{1+\sum_{j\neq k^\star}\exp(-\delta_j)}. \tag{22}$$

Since $m(x;\theta) = \min_{j\neq k^\star}\delta_j$, we have $\exp(-\delta_j) \le \exp(-m)$ for all $j \neq k^\star$, yielding

$$\sum_{j\neq k^\star}\exp(-\delta_j) \le (K-1)\exp(-m), \tag{23}$$

which proves the lower bound. Conversely, since there exists at least one $j_0 \neq k^\star$ such that $\delta_{j_0} = m$, we obtain

$$\sum_{j\neq k^\star}\exp(-\delta_j) \ge \exp(-m), \tag{24}$$

which implies the upper bound. □

For notational simplicity, define $\sigma(t) \triangleq 1/(1+\exp(-t))$ and $g(t) \triangleq 1/(1+(K-1)\exp(-t))$.

By Lemma 1,

$$\mathbb{E}_{x_i\sim\mathcal{D}_{\text{ID}}}[S(x_i;\theta)] \le \mathbb{E}[\sigma(m_i)]. \tag{25}$$

The sigmoid function $\sigma(t)$ satisfies

$$\sigma''(t) = \sigma(t)\big(1-\sigma(t)\big)\big(1-2\sigma(t)\big) \le 0, \quad \forall\, t \ge 0, \tag{26}$$

and is therefore concave on $[0,\infty)$. In practice, ID samples are typically well classified with positive margins $m_i \ge 0$ with high probability. Under this mild assumption, Jensen's inequality yields

$$\mathbb{E}[\sigma(m_i)] \le \sigma(\mathbb{E}[m_i]) = \frac{1}{1+\exp(-\mu_i)}. \tag{27}$$

Any deviation caused by rare negative-margin ID samples is absorbed into a higher-order remainder term. Again by Lemma 1,

$$\mathbb{E}_{x_o\sim\mathcal{D}_{\text{OOD}}}[S(x_o;\theta)] \geq \mathbb{E}[g(m_o)]. \tag{28}$$

We perform a second-order Taylor expansion of $g(m_o)$ around μ_o. Let $\varepsilon = m_o - \mu_o$, then $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\varepsilon^2] = \nu_o^2$. By Taylor's theorem,

$$g(\mu_o+\varepsilon) = g(\mu_o) + g'(\mu_o)\varepsilon + \frac{1}{2}g''(\mu_o)\varepsilon^2 + \frac{1}{6}g^{(3)}(\xi)\varepsilon^3, \tag{29}$$

where ξ lies between μ_o and $\mu_o+\varepsilon$. Taking expectation gives

$$\mathbb{E}[g(m_o)] = g(\mu_o) + \frac{1}{2}g''(\mu_o)\nu_o^2 + \mathcal{R}_3, \tag{30}$$

where $\mathcal{R}_3 = \frac{1}{6}\mathbb{E}[g^{(3)}(\xi)\varepsilon^3]$ collects third- and higher-order terms.

A direct calculation yields

$$g'(t) = \frac{(K-1)\exp(-t)}{(1+(K-1)\exp(-t))^2}, \tag{31}$$

and

$$g''(t) = \frac{(K-1)(K-1-\exp(t))\exp(-t)}{(1+(K-1)\exp(-t))^3}. \tag{32}$$

In the over-softening regime induced by uniform OE supervision, $\mu_o \leq \log(K-1)$ typically holds, implying $g''(\mu_o) \geq 0$. Using the bound $g''(\mu_o) \leq g'(\mu_o)$, we obtain

$$\begin{aligned}\mathbb{E}[g(m_o)] \geq\ & \frac{1}{1+(K-1)\exp(-\mu_o)} \\ & + \frac{(K-1)\exp(-\mu_o)}{2(1+(K-1)\exp(-\mu_o))^2}\nu_o^2 + \widetilde{\mathcal{R}}.\end{aligned} \tag{33}$$

where $\widetilde{\mathcal{R}}$ absorbs the discrepancy between g'' and g' as well as the third-order remainder.

Combining the ID upper bound and the OOD lower bound, we obtain

$$\begin{aligned}\Delta(\theta) \leq\ & \frac{1}{1+\exp(-\mu_i)} - \frac{1}{1+(K-1)\exp(-\mu_o)} \\ & - \frac{(K-1)\exp(-\mu_o)}{2(1+(K-1)\exp(-\mu_o))^2}\nu_o^2 + \mathcal{R},\end{aligned} \tag{34}$$

where $\mathcal{R}$ collects all higher-order and approximation terms. This completes the proof.

Proof of Theorem 2

For an outlier sample x_o, we minimize $\mathcal{L} = \text{KL}(q_T \| s(z))$, where $q_T = s(z/T)$ is the temperature-smoothed target. The gradient with respect to z_k is $\frac{\partial \mathcal{L}}{\partial z_k} = p_k - q_{T,k}$. Let $a = \arg\max_k z_k$ and $b = \arg\max_{j\neq a} z_j$. The update for the logit margin $m(z) = z_a - z_b$ is given by:

$$m(z^+) = (z_a - \eta(p_a - q_{T,a})) - (z_b - \eta(p_b - q_{T,b})) \tag{35}$$

$$= m(z) - \eta[(p_a - p_b) - (q_{T,a} - q_{T,b})]. \tag{36}$$

The term $p_a - p_b$ is expanded as:

$$\begin{aligned}p_a - p_b &= \frac{e^{z_a} - e^{z_b}}{\sum_{k=1}^K e^{z_k}} \\ &= \frac{e^{z_a - z_b} - 1}{e^{z_a - z_b} + 1 + \sum_{k\neq a,b} e^{z_k - z_b}} \\ &= \frac{e^{m(z)} - 1}{e^{m(z)} + 1 + R(z,b)}.\end{aligned} \tag{37}$$

where $R(z,b) = \sum_{k\neq a,b}\exp(z_k - z_b)$. Similarly, for the target distribution with temperature T:

$$q_{T,a} - q_{T,b} = \frac{e^{z_a/T} - e^{z_b/T}}{\sum_{k=1}^K e^{z_k/T}} = \frac{e^{m(z)/T} - 1}{e^{m(z)/T} + 1 + R(z/T,b)}. \tag{38}$$

Substituting these terms back into the update equation yields:

$$m(z^+) = m(z) - \eta\left[\frac{e^{m(z)} - 1}{e^{m(z)} + 1 + R(z,b)} - \frac{e^{m(z)/T} - 1}{e^{m(z)/T} + 1 + R(z/T,b)}\right]. \tag{39}$$

This completes the proof.

Proof of Corollary 1

Define the function $h(x,R) = \frac{e^x - 1}{e^x + 1 + R}$. The margin contraction is given by:

$$\Delta m_T = \eta\left[h(m(z),R(z,b)) - h\left(\frac{m(z)}{T}, R\left(\frac{z}{T},b\right)\right)\right]. \tag{40}$$

In standard uniform OE ($T\to\infty$), $m(z)/T \to 0$ and $z_k/T \to 0$, leading to $h(0,K-2) = 0$. Thus, the maximum contraction is $\Delta m_\infty = \eta h(m(z),R(z,b))$. To analyze the monotonicity, we compute the partial derivative of $h(x,R)$ with respect to x:

$$\frac{\partial h}{\partial x} = \frac{e^x(e^x+1+R) - e^x(e^x-1)}{(e^x+1+R)^2} = \frac{e^x(R+2)}{(e^x+1+R)^2} > 0. \tag{41}$$

Since h is strictly increasing in x, and for any $T > 1$, $0 < \frac{m(z)}{T} < m(z)$, we have:

$$0 < h\left(\frac{m(z)}{T}, R\left(\frac{z}{T},b\right)\right) < h(m(z),R(z,b)). \tag{42}$$

Consequently, $\Delta m_T = \Delta m_\infty - \eta h(m(z)/T, R(z/T,b)) < \Delta m_\infty$. This proves that a finite temperature $T^\star \in (1,\infty)$ strictly reduces the magnitude of margin contraction compared to uniform OE.

Generalization Bound

Preliminary

For clarity and ease of reference, we summarize the theorems and corollaries presented in this paper.

Theorem 3. Define the true relationship between OOD sample $\boldsymbol{x}$ and ID categories as $p^\star(\boldsymbol{x})$, where $p^\star \sim \mathcal{T}$. And define the assigned pseudo-labels as $q(\boldsymbol{x}) = \text{softmax}(f(\boldsymbol{x})/T)$ for some temperature $T > 1$, $q(\boldsymbol{x}) \sim \mathcal{Q}$. Then, the expected diversity between the pseudo-labels and the true relationship satisfies the following upper bound:

$$\mathbb{E}\left[\text{KL}\left(\mathcal{T}\,\|\,\mathcal{Q}\right)\right] \leq \log K - \mathbb{E}\left[H(p^\star(\boldsymbol{x}))\right] - \frac{C_1'}{T} + \frac{C_2'}{T^2} + O(\frac{1}{T^3}), \tag{43}$$

where $C_1' = \mathbb{E}_{\boldsymbol{x}}\left[\sum_{i=1}^K p_i(\boldsymbol{x})f_i(\boldsymbol{x}) - \mu(\boldsymbol{x})\right]$, $C_2' = \frac{1}{2}\mathbb{E}_{\boldsymbol{x}}\left[\text{Var}(f(\boldsymbol{x}))\right]$, with $\mu(\boldsymbol{x}) = \frac{1}{K}\sum_{i=1}^K f_i(\boldsymbol{x})$ and $\text{Var}(f(\boldsymbol{x})) = \frac{1}{K}\sum_{i=1}^K f_i(\boldsymbol{x})^2 - \mu(\boldsymbol{x})^2$. And $p(\boldsymbol{x}) = \text{softmax}(f(\boldsymbol{x}))$.

Corollary 2 (Optimal Temperature)**.** Under the conditions of Theorem 3, the expected KL divergence achieves its minimum at a finite temperature $T^\star = \frac{2C_2'}{C_1'}$, where C_1' and C_2' are defined as in Theorem 3.

Corollary 3 (Optimal Temperature is Strictly Better than Infinite Temperature)**.** Under the conditions of Theorem 3, the KL divergence evaluated at the optimal finite temperature $T^\star = \frac{2C_2'}{C_1'}$ is strictly smaller than the KL divergence as $T\to\infty$. Formally, we have

$$\mathbb{E}\left[\text{KL}\left(\mathcal{T}\,\|\,\mathcal{Q}\right)\right]\Big|_{T=T^\star} < \lim_{T\to\infty}\mathbb{E}\left[\text{KL}\left(\mathcal{T}\,\|\,\mathcal{Q}\right)\right]. \tag{44}$$

Proof of Theorem 3

Let $f(x) \in \mathbb{R}^K$ denote the logit output of a model over K classes. To model predictive uncertainty, particularly *epistemic uncertainty*, we consider perturbations in the logit space, following prior works on approximate Bayesian inference via stochastic ensembles and MC Dropout [63]. Specifically, we define the predictive distribution under logit noise as $\tilde{p}(x) := \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2 I)}[\text{softmax}(f(x)+\epsilon)]$, where $\epsilon \sim \mathcal{N}(0,\sigma^2 I)$ represents Gaussian perturbations that approximate the variability induced by the posterior over model parameters. This formulation approximates the Bayesian predictive distribution $p(y \mid x, \mathcal{D})$, marginalized over the posterior $p(\theta \mid \mathcal{D})$, under the assumption that variations in $f_\theta(x)$ can be locally modeled as additive Gaussian noise in the logit space. In practice, we approximate $p^\star(x)$ with a single Monte Carlo sample:

$$p^\star(x) := \text{softmax}(f(x)+\epsilon), \quad \epsilon \sim \mathcal{N}(0,\sigma^2 I), \tag{45}$$

where $p^\star(x)$ serves as a stochastic proxy of the true predictive distribution. Additionally, we define a temperature-scaled softmax distribution:

$$q(x) := \text{softmax}\left(\frac{f(x)}{T}\right), \quad T > 1, \tag{46}$$

where the temperature parameter T controls the entropy of the distribution. A higher temperature yields a softer output with higher uncertainty, which is particularly useful for pseudo-labeling and calibration.

We analyze the Kullback–Leibler divergence [36]:

$$\text{Gap}(T) := \text{KL}(p^\star(x)\|q(x)). \tag{47}$$

By definition of KL divergence:

$$\text{Gap}(T) = \sum_{i=1}^K p_i^\star(x)\log\left(\frac{p_i^\star(x)}{q_i(x)}\right) \tag{48}$$

$$= -H(p^\star(x)) - \sum_{i=1}^K p_i^\star(x)\log q_i(x), \tag{49}$$

where $H(p^\star(x)) = -\sum_{i=1}^K p_i^\star(x)\log p_i^\star(x)$ denotes the entropy.

Since

$$\log q_i(x) = \frac{f_i(x)}{T} - \log\left(\sum_{j=1}^K e^{f_j(x)/T}\right), \tag{50}$$

we obtain:

$$\sum_{i=1}^K p_i^\star(x)\log q_i(x) = \frac{1}{T}\sum_{i=1}^K p_i^\star(x) f_i(x) - \log\left(\sum_{j=1}^K e^{f_j(x)/T}\right). \tag{51}$$

Substituting back, we get:

$$\text{Gap}(T) = -H(p^\star(x)) - \frac{1}{T}\sum_{i=1}^K p_i^\star(x) f_i(x) + \log\left(\sum_{j=1}^K e^{f_j(x)/T}\right). \tag{52}$$

Assume that the noise magnitude $\|\epsilon\|$ is sufficiently small. Expanding $p_i^\star(x)$ via first-order Taylor approximation:

$$p_i^\star(x) = p_i(x) + \sum_{j=1}^K \frac{\partial p_i(x)}{\partial f_j(x)}\epsilon_j + o(\|\epsilon\|), \tag{53}$$

where $\frac{\partial p_i}{\partial f_j} = p_i(x)(\delta_{ij} - p_j(x))$.

Taking expectation with respect to ϵ and using $\mathbb{E}[\epsilon_j] = 0$, we obtain:

$$\mathbb{E}[p_i^\star(x)] = p_i(x) + O(\sigma^2), \tag{54}$$

$$\mathbb{E}\left[\sum_i p_i^\star(x) f_i(x)\right] = \sum_i p_i(x) f_i(x) + O(\sigma^2), \tag{55}$$

$$\mathbb{E}[H(p^\star(x))] = H(p(x)) + O(\sigma^2). \tag{56}$$

Next, define the log-partition function $Z_T(x) := \sum_{j=1}^K e^{f_j(x)/T}$. Using a second-order Taylor expansion:

$$\log Z_T(x) = \log K + \frac{\mu(x)}{T} + \frac{1}{2T^2}\text{Var}(f(x)) + O\left(\frac{1}{T^3}\right), \tag{57}$$

where $\mu(x) := \frac{1}{K}\sum_{j=1}^K f_j(x)$ and $\text{Var}(f(x)) := \frac{1}{K}\sum_{j=1}^K f_j(x)^2 - \mu(x)^2$.

Substituting (54)–(57) into (52), we obtain:

$$\mathbb{E}[\text{Gap}(T)] = -\mathbb{E}[H(p^\star(x))] - \frac{1}{T}\mathbb{E}\left[\sum_i p_i(x) f_i(x)\right] + \log K + \frac{\mathbb{E}[\mu(x)]}{T} + \frac{1}{2T^2}\mathbb{E}[\text{Var}(f(x))] + O\left(\frac{1}{T^3}\right). \tag{58}$$

Define

$$C_1' := \mathbb{E}\left[\sum_i p_i(x) f_i(x)\right] - \mathbb{E}[\mu(x)], \tag{59}$$

and

$$C_2' := \frac{1}{2}\mathbb{E}[\text{Var}(f(x))]. \tag{60}$$

Then the KL divergence admits the following upper bound:

$$\mathbb{E}[\text{Gap}(T)] \le \log K - \mathbb{E}[H(p^\star(x))] - \frac{C_1'}{T} + \frac{C_2'}{T^2} + O\left(\frac{1}{T^3}\right), \tag{61}$$

where $C_1', C_2' > 0$.

Proof of Corollary 2

The leading terms of the bound in (61) define a function $g(T) := -\frac{C_1'}{T} + \frac{C_2'}{T^2}$. Setting the derivative $g'(T) = 0$ gives $\frac{C_1'}{T^2} = \frac{2C_2'}{T^3}$, implying the optimal temperature is $T^\star = \frac{2C_2'}{C_1'}$.

Proof of Corollary 3

Substituting $T^\star = \frac{2C_2'}{C_1'}$ into the dominant terms yields:

$$-\frac{C_1'}{T^\star} + \frac{C_2'}{(T^\star)^2} = -\frac{(C_1')^2}{4C_2'} < 0, \tag{62}$$

where the inequality follows from the positivity of C_1' and C_2'. Thus, the KL divergence at $T = T^\star$ is strictly smaller than that at $T \to \infty$.

Fine-grained results

In addition to comparing with methods that incorporate outliers during the training of the K classifier [5, 14, 16, 27], we also include comparisons with approaches that train an additional binary classifier specifically for OOD detection, such as NPOS [64], VOS [65]. The corresponding results are reported in Table 8, Table 9, Table 10, Table 11, Table 12, and Table 13.

Table 8 Fine-grained results (FPR95 ↓) on the CIFAR-10 benchmark.

Method	CIFAR100	TIN	MNIST	SVHN	Textures	Places365
OE	36.71 ± 2.06	2.97 ± 1.17	24.67 ± 2.55	1.25 ± 0.36	12.07 ± 2.14	14.53 ± 2.80
MixOE	58.29 ± 8.25	44.62 ± 7.57	38.28 ± 13.40	20.36 ± 3.99	33.19 ± 4.28	43.54 ± 4.95
DOE	28.27 ± 0.84	12.50 ± 0.95	35.70 ± 3.33	2.55 ± 0.78	10.35 ± 1.50	13.77 ± 0.47
DAL	30.07 ± 1.56	11.76 ± 1.09	55.11 ± 8.52	3.24 ± 1.09	13.00 ± 1.13	14.26 ± 0.86
NPOS	35.71 ± 1.17	29.57 ± 0.24	22.96 ± 2.27	6.41 ± 0.19	20.80 ± 2.19	32.19 ± 0.98
VOS	61.57 ± 3.24	52.49 ± 0.73	35.92 ± 11.38	31.50 ± 8.38	46.53 ± 5.24	47.78 ± 3.62
AOE-At	33.54 ± 2.24	3.21 ± 1.30	26.15 ± 3.19	0.96 ± 0.26	9.05 ± 1.29	11.35 ± 1.87
AOE-Jt	33.60 ± 0.24	1.27 ± 0.16	21.76 ± 5.00	0.50 ± 0.16	9.48 ± 3.42	10.72 ± 2.39

Table 9 Fine-grained results (AUROC ↑) on the CIFAR-10 benchmark.

Method	CIFAR100	TIN	MNIST	SVHN	Textures	Places365
OE	90.54 ± 0.53	99.11 ± 0.34	90.22 ± 1.31	99.60 ± 0.14	97.58 ± 0.27	96.58 ± 0.70
MixOE	87.47 ± 0.97	90.00 ± 0.73	91.66 ± 2.21	93.82 ± 1.27	91.84 ± 0.51	90.38 ± 0.55
DOE	92.63 ± 0.21	97.04 ± 0.27	85.16 ± 2.20	99.20 ± 0.09	97.77 ± 0.31	96.57 ± 0.18
DAL	91.89 ± 0.46	96.95 ± 0.38	75.35 ± 5.05	99.10 ± 0.19	97.04 ± 0.18	96.18 ± 0.20
NPOS	88.57 ± 0.36	90.99 ± 0.35	92.64 ± 1.59	98.88 ± 0.09	94.44 ± 0.90	90.32 ± 0.74
VOS	86.57 ± 0.57	88.84 ± 0.48	91.56 ± 2.37	92.18 ± 1.65	89.68 ± 1.32	89.90 ± 0.66
AOE-At	91.34 ± 0.48	99.08 ± 0.30	89.66 ± 2.02	99.68 ± 0.07	98.19 ± 0.19	97.46 ± 0.43
AOE-Jt	91.68 ± 0.13	99.60 ± 0.07	92.75 ± 1.85	99.84 ± 0.05	98.07 ± 0.63	97.50 ± 0.74

Table 10 Fine-grained results (FPR95 ↓) on the CIFAR-100 benchmark.

Method	CIFAR10	TIN	MNIST	SVHN	Textures	Places365
OE	61.26 ± 0.22	0.21 ± 0.01	53.31 ± 9.91	51.84 ± 3.45	55.83 ± 1.82	58.30 ± 0.72
MixOE	61.12 ± 1.08	49.32 ± 0.36	59.49 ± 7.74	73.09 ± 4.00	66.04 ± 0.98	56.93 ± 0.78
DOE	63.85 ± 0.50	11.83 ± 1.60	57.97 ± 4.42	20.27 ± 1.01	50.29 ± 2.73	52.97 ± 1.56
DAL	65.70 ± 0.96	4.86 ± 1.57	65.29 ± 4.51	12.41 ± 2.75	55.04 ± 1.44	56.59 ± 1.70
NPOS	72.50 ± 2.50	54.21 ± 1.09	66.98 ± 4.61	30.67 ± 0.74	47.39 ± 1.10	59.47 ± 0.21
VOS	59.23 ± 0.59	51.89 ± 1.01	48.56 ± 2.00	47.23 ± 3.04	62.55 ± 1.04	56.44 ± 0.36
AOE-At	60.74 ± 0.22	0.38 ± 0.13	46.91 ± 3.26	27.25 ± 13.06	52.17 ± 1.46	56.53 ± 1.12
AOE-Jt	59.91 ± 0.47	0.27 ± 0.05	51.00 ± 5.67	24.72 ± 9.80	52.02 ± 2.14	56.74 ± 0.38

Table 11 Fine-grained results (AUROC ↑) on the CIFAR-100 benchmark.

Method	CIFAR10	TIN	MNIST	SVHN	Textures	Places365
OE	76.70 ± 0.19	99.89 ± 0.02	80.68 ± 5.82	84.37 ± 1.34	82.18 ± 0.68	78.39 ± 0.41
MixOE	78.17 ± 0.29	83.73 ± 0.12	76.06 ± 5.52	72.28 ± 0.81	77.34 ± 0.91	79.92 ± 0.30
DOE	75.47 ± 0.55	97.76 ± 0.28	72.54 ± 4.38	95.73 ± 0.42	86.34 ± 1.28	82.58 ± 0.91
DAL	73.08 ± 0.19	98.86 ± 0.32	67.91 ± 2.00	97.12 ± 0.56	84.04 ± 0.62	80.60 ± 0.94
NPOS	75.37 ± 0.58	81.32 ± 0.17	73.26 ± 5.32	92.43 ± 0.29	85.55 ± 0.40	77.92 ± 0.37
VOS	79.14 ± 0.41	82.73 ± 0.20	82.29 ± 1.51	84.23 ± 1.33	78.41 ± 0.78	80.34 ± 0.03
AOE-At	76.67 ± 0.12	99.81 ± 0.04	85.52 ± 2.32	93.82 ± 3.59	84.65 ± 1.31	79.88 ± 0.97
AOE-Jt	76.76 ± 0.22	99.88 ± 0.01	84.41 ± 2.69	95.76 ± 1.84	84.29 ± 0.71	79.70 ± 0.25

Table 12 Fine-grained results (FPR95 ↓) on the ImageNet-200 benchmark.

Method	SSB-hard	NINCO	iNaturalist	Textures	OpenImage-O
OE	64.20 ± 0.45	40.90 ± 0.66	28.93 ± 0.58	41.82 ± 1.09	34.78 ± 0.24
MixOE	67.94 ± 0.43	47.95 ± 0.63	29.97 ± 0.20	50.23 ± 1.21	40.17 ± 0.67
DOE	63.70 ± 0.04	44.59 ± 0.99	29.13 ± 3.75	44.87 ± 1.93	38.81 ± 3.35
DAL	61.68 ± 0.12	42.03 ± 0.71	27.58 ± 0.45	44.84 ± 0.40	34.86 ± 0.31
NPOS	74.29 ± 0.52	49.89 ± 0.49	20.01 ± 0.69	16.87 ± 0.19	28.40 ± 0.28
VOS	69.93 ± 0.47	49.85 ± 0.71	25.53 ± 1.36	39.74 ± 1.17	36.77 ± 0.94
AOE-At	62.70 ± 0.13	38.36 ± 0.03	28.58 ± 0.74	41.97 ± 0.37	33.27 ± 0.22
AOE-Jt	62.62 ± 0.41	39.04 ± 0.44	26.61 ± 0.85	40.57 ± 0.39	32.06 ± 0.61

Table 13 Fine-grained results (AUROC ↑) on the ImageNet-200 benchmark.

Method	SSB-hard	NINCO	iNaturalist	Textures	OpenImage-O
OE	82.14 ± 0.17	86.89 ± 0.25	89.08 ± 0.18	87.33 ± 0.22	88.22 ± 0.19
MixOE	80.15 ± 0.21	84.99 ± 0.28	90.94 ± 0.17	87.02 ± 0.22	87.22 ± 0.04
DOE	80.64 ± 0.71	85.81 ± 0.54	90.64 ± 1.21	87.04 ± 1.74	87.05 ± 1.41
DAL	83.34 ± 0.12	86.99 ± 0.04	90.00 ± 0.12	87.18 ± 0.19	88.46 ± 0.04
NPOS	74.29 ± 0.42	84.50 ± 0.40	94.81 ± 0.15	96.97 ± 0.04	91.69 ± 0.04
VOS	79.68 ± 0.19	85.35 ± 0.10	92.77 ± 0.54	90.95 ± 0.20	89.28 ± 0.15
AOE-At	83.46 ± 0.02	87.77 ± 0.15	89.82 ± 0.37	87.83 ± 0.02	89.19 ± 0.06
AOE-Jt	83.49 ± 0.15	87.65 ± 0.02	90.36 ± 0.30	88.15 ± 0.07	89.21 ± 0.04

References

[1] Hendrycks D, Gimpel K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations. 2017

[2] Yang J, Zhou K, Li Y, Liu Z. Generalized out-of-distribution detection: A survey. International Journal of Computer Vision, 2024, 132(12): 5635–5662

[3] Yang Y, Guo J, Li G, Li L, Li W, Yang J. Alignment efficient image-sentence retrieval considering transferable cross-modal representation learning. Frontiers of Computer Science, 2024, 18(3): 181335

[4] Guo L, Jia L, Shao J, Li Y. Robust semi-supervised learning in open environments. Frontiers of Computer Science, 2025, 19(8): 198345

[5] Wang Q, Ye J, Liu F, Dai Q, Kalander M, Liu T, Hao J, Han B. Out-of-distribution detection with implicit outlier transformation. In: International Conference on Learning Representations. 2023

[6] Lee K, Lee H, Lee K, Shin J. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In: International Conference on Learning Representations. 2018

[7] Liu W, Wang X, Owens J D, Li Y. Energy-based out-of-distribution detection. In: Neural Information Processing Systems. 2020, 21464–21475

[8] Djurisic A, Bozanic N, Ashok A, Liu R. Extremely simple activation shaping for out-of-distribution detection. In: International Conference on Learning Representations. 2023

[9] Sun Y, Li Y. DICE: leveraging sparsification for out-of-distribution detection. In: European Conference on Computer Vision. 2022, 691–708

[10] Yang Y, Jiang N J, Xu Y, Zhan D. Robust semi-supervised learning by wisely leveraging open-set data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 8334–8347

[11] Xu H, Yang Y. Shaping parameter contribution patterns for out-of-distribution detection. In: Association for the Advancement of Artificial Intelligence. 2026, 27242–27250

[12] Wei H, Xie R, Cheng H, Feng L, An B, Li Y. Mitigating neural network overconfidence with logit normalization. In: International Conference on Machine Learning. 2022, 23631–23644

[13] Pei S. Image background serves as good proxy for out-of-distribution data. In: International Conference on Learning Representations. 2024

[14] Hendrycks D, Mazeika M, Dietterich T G. Deep anomaly detection with outlier exposure. In: International Conference on Learning Representations. 2019

[15] Jiang W, Cheng H, Chen M, Wang C, Wei H. DOS: diverse outlier sampling for out-of-distribution detection. In: International Conference on Learning Representations. 2024

[16] Wang Q, Fang Z, Zhang Y, Liu F, Li Y, Han B. Learning to augment distributions for out-of-distribution detection. In: Neural Information Processing Systems. 2023

[17] Wang Y, Qiao P, Liu C, Song G, Zheng X, Chen J. Out-of-distributed semantic pruning for robust semi-supervised learning. In: Computer Vision and Pattern Recognition. 2023, 23849–23858

[18] Wallin E, Svensson L, Kahl F, Hammarstrand L. Prosub: Probabilistic open-set semi-supervised learning with subspace-based out-of-distribution detection. In: European Conference on Computer Vision. 2024, 129–147

[19] Chen Q, Ding H. Dual energy-based model with open-world uncertainty estimation for out-of-distribution detection. In: Computer Vision and Pattern Recognition. 2025, 25728–25737

[20] Li S, Zhao S, Cao Z, Huang S, Chen S. Robust domain adaptation with noisy and shifted label distribution. Frontiers of Computer Science, 2025, 19(3): 193310

[21] Huang R, Geng A, Li Y. On the importance of gradients for detecting distributional shifts in the wild. In: Neural Information Processing Systems. 2021, 677–689

[22] Sun Y, Ming Y, Zhu X, Li Y. Out-of-distribution detection with deep nearest neighbors. In: International Conference on Machine Learning. 2022, 20827–20840

[23] Regmi S, Panthi B, Dotel S, Gyawali P K, Stoyanov D, Bhattarai B. T2fnorm: Train-time feature normalization for OOD detection in image classification. In: Computer Vision and Pattern Recognition. 2024, 153–162

[24] Ghosal S S, Sun Y, Li Y. How to overcome curse-of-dimensionality for out-of-distribution detection? In: Association for the Advancement of Artificial Intelligence. 2024, 19849–19857

[25] Zhu J, Li H, Yao J, Liu T, Xu J, Han B. Unleashing mask: Explore the intrinsic out-of-distribution detection capability. In: International Conference on Machine Learning. 2023, 43068–43104

[26] Yang Y, Xu H. Strengthen out-of-distribution detection capability with progressive self-knowledge distillation. In: International Conference on Machine Learning. 2025

[27] Zhang J, Inkawhich N, Linderman R, Chen Y, Li H. Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments. In: Winter Conference on Applications of Computer Vision. 2023, 5520–5529

[28] Miao W, Pang G, Bai X, Li T, Zheng J. Out-of-distribution detection in long-tailed recognition with calibrated outlier class learning. In: Association for the Advancement of Artificial Intelligence. 2024, 4216–4224

[29] Guo C, Pleiss G, Sun Y, Weinberger K Q. On calibration of modern neural networks. In: International Conference on Machine Learning. 2017, 1321–1330

[30] Hinton G E, Vinyals O, Dean J. Distilling the knowledge in a neural network. CoRR, 2015

[31] Yu Y, Bates S, Ma Y, Jordan M I. Robust calibration with multi-domain temperature scaling. In: Neural Information Processing Systems. 2022, 27510–27523

[32] Hwang S, Kim M, Whang S E. T-CIL: temperature scaling using adversarial perturbation for calibration in class-incremental learning. In: Computer Vision and Pattern Recognition. 2025, 15339–15348

[33] Chen T, Kornblith S, Norouzi M, Hinton G E. A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. 2020, 1597–1607

[34] Santos d C N, Xiang B, Zhou B. Classifying relations by ranking with convolutional neural networks. In: Annual Meeting of the Association for Computational Linguistics. 2015, 626–634

[35] Cassotti P, Pascale S D, Tahmasebi N. Using synchronic definitions and semantic relations to classify semantic change types. In: Annual Meeting of the Association for Computational Linguistics. 2024, 4539–4553

[36] Ji S, Zhang Z, Ying S, Wang L, Zhao X, Gao Y. Kullback-leibler divergence metric learning. IEEE Transactions on Cybernetics, 2022, 52(4): 2047–2058

[37] Dang H, Huu T T, Nguyen T M, Ho N. Neural collapse for cross-entropy class-imbalanced learning with unconstrained relu features model. In: International Conference on Machine Learning. 2024, 10017–10040

[38] Parvasi S P, Musavi M, Torabi S A, Yap W Y, Lam J S L. Enhancing transportation service management with bi-level optimization: Competitive pricing and hub location. European Journal of Operational Research, 2026, 328(3): 815–831

[39] Deng Z, Li D, Peng X, Song Y, Xiang T. BORT2: bi-level optimization for robust target training in multi-source domain adaptation. Pattern Recognition, 2026, 172: 112367

[40] Zhang J, Yang J, Wang P, Wang H, Lin Y, Zhang H, Sun Y, Du X, Li Y, Liu Z, Chen Y, Li H. Openood v1.5: Enhanced benchmark for out-of-distribution detection. Data-centric Machine Learning Research, 2024, 2: (3):1–32

[41] Krizhevsky A, Hinton G, others . Learning multiple layers of features from tiny images. 2009

[42] Abai Z, Rajmalwar N. Densenet models for tiny imagenet classification. CoRR, 2019

[43] Deng L. The MNIST database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine, 2012, 29(6): 141–142

[44] Netzer Y, Wang T, Coates A, Bissacco A, Wu B, Ng A Y, others . Reading digits in natural images with unsupervised feature learning. In: Neural Information Processing Systems. 2011, 4
[45] Gool L V, Vandermeulen D, Kalberer G A, Tuytelaars T, Zalesny A. Modeling shapes and textures from images: new frontiers. In: 3D Data Processing, Visualization and Transmission. 2002, 286–295
[46] Zhou B, Lapedriza A, Khosla A, Oliva A, Torralba A. Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(6): 1452–1464
[47] Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition. 2009, 248–255
[48] Vaze S, Han K, Vedaldi A, Zisserman A. Open-set recognition: A good closed-set classifier is all you need. In: International Conference on Learning Representations. 2022
[49] Bitterwolf J, Müller M, Hein M. In or out? fixing imagenet out-of-distribution detection evaluation. In: International Conference on Machine Learning. 2023, 2471–2506
[50] Horn G V, Aodha O M, Song Y, Cui Y, Sun C, Shepard A, Adam H, Perona P, Belongie S J. The inaturalist species classification and detection dataset. In: Computer Vision and Pattern Recognition. 2018, 8769–8778
[51] Wang H, Li Z, Feng L, Zhang W. Vim: Out-of-distribution with virtual-logit matching. In: Computer Vision and Pattern Recognition. 2022, 4911–4920
[52] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition. 2016, 770–778
[53] Loshchilov I, Hutter F. SGDR: stochastic gradient descent with warm restarts. In: International Conference on Learning Representations. 2017
[54] Liang S, Li Y, Srikant R. Enhancing the reliability of out-of-distribution image detection in neural networks. In: International Conference on Learning Representations. 2018
[55] Sun Y, Guo C, Li Y. React: Out-of-distribution detection with rectified activations. In: Neural Information Processing Systems. 2021, 144–157
[56] Xu K, Chen R, Franchi G, Yao A. Scaling for training time and post-hoc out-of-distribution detection enhancement. In: International Conference on Learning Representations. 2024
[57] Zhang J, Fu Q, Chen X, Du L, Li Z, Wang G, Liu X, Han S, Zhang D. Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In: International Conference on Learning Representations. 2023
[58] Liu X, Lochman Y, Zach C. GEN: pushing the limits of softmax-based out-of-distribution detection. In: Computer Vision and Pattern Recognition. 2023, 23946–23955
[59] Song Y, Wang W, Sebe N. Rankfeat&rankweight: Rank-1 feature weight removal for out-of-distribution detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2025, 47(4): 2505–2519
[60] Tamang L D, Bouadjenek M R, Dazeley R, Aryal S. Margin-bounded confidence scores for out-of-distribution detection. In: International Conference on Data Mining. 2024, 1–10
[61] Ming Y, Fan Y, Li Y. POEM: out-of-distribution detection with posterior sampling. In: International Conference on Machine Learning. 2022, 15650–15665
[62] Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: International Conference on Computer Vision. 2017, 618–626
[63] Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Neural Information Processing Systems. 2017, 6402–6413
[64] Tao L, Du X, Zhu J, Li Y. Non-parametric outlier synthesis. In: International Conference on Learning Representations. 2023
[65] Du X, Wang Z, Cai M, Li Y. VOS: learning what you don't know by virtual outlier synthesis. In: International Conference on Learning Representations. 2022

Fengqiang Wan is currently working toward the Ph.D. degree at the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests mainly lie in deep learning and data mining.

Qingyuan Jiang received the BSc and the PhD degrees in computer science from Nanjing University, China. He has published more than ten papers in leading international journals/conferences. He serves as a PC/Reviewer in leading conferences such as IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), IEEE Transactions on Image Processing (TIP), NeurIPS, ICML, ICLR, AISTATS, AAAI, IJCAI, etc. He is currently an associate professor at Nanjing University of Science and Technology. His research interests are in machine learning, multimodal learning, and information retrieval.

Fu Shen received the MSc degrees in computer science from Southeast University, China. He has published more than ten papers in SCI and Chinese core journals. He serves as an associate researcher at the Nanjing Institute of Agricultural Mechanization,Ministry of Agriculture and Rural Affairs. His research interests encompass agricultural big data model analysis, decision-making of farmland environmental data, and autonomous homework navigation trajectory algorithm.

Yang Yang received his Ph.D. in Computer Science from Nanjing University, China, in 2019. In the same year, he joined Nanjing University of Science and Technology, where he is currently a Professor in the School of Computer Science and Engineering. His research focuses on machine learning and data mining, with particular interests in heterogeneous learning, model reuse, and incremental mining. He has published extensively in leading journals and conferences, including TKDE, ACM TOIS, TKDD, SIGKDD, SIGIR, WWW, IJCAI, and AAAI. He received the Best Paper Award at ACML 2017. He also serves as a program committee member for major conferences such as IJCAI, AAAI, ICML, and NeurIPS.