南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Robust Semi-supervised Learning by Wisely Leveraging Open-set Data

Yang Yang, *Member, IEEE,* Nan Jiang, Yi Xu, and De-Chuan Zhan

TPAMI 2024

Introduction

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

(a) Semi-supervised Learning(SSL)

- Leverages the ubiquitous **unlabeled data** to break the limitation of supervised learning (SL) caused by the huge human and financial costs in obtaining labeled data
- Traditional SSL typically **makes the assumption** that labeled data and unlabeled data **share the same class space**.However, in real applications, the unlabeled training dataset may **contain the data from classes unseen in the labeled**.

(b) Open-set Semi-supervised Learning(OSSL)

- Usually apply an **OOD detection module** in addition to the **traditional ID classifier**, for the purpose of acquiring the capability of **differentiating OOD data from ID data**.
- Typically, all open-set (OS) data are involved in these OSSL models' training, which may **comprise both friendly data and unfriendly data.**

Introduction

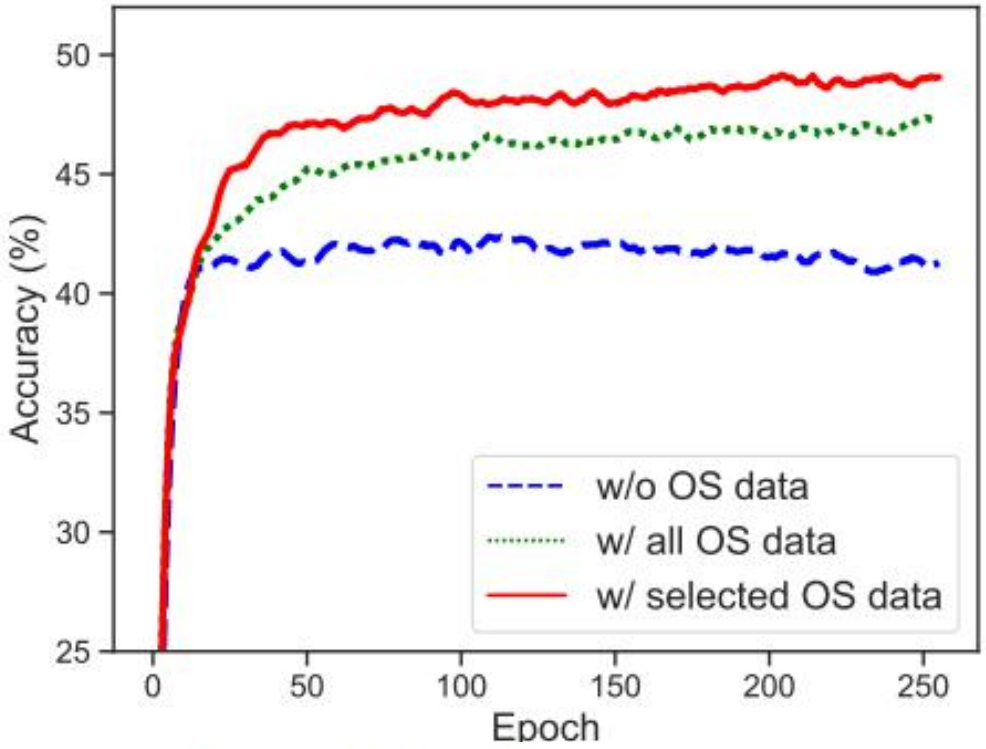


Fig. 1: An example of models' performance (testing accuracy on ID classification) with different strategies of using the open-set data (OS data) illustrates the effectiveness of selectively leveraging OS data during the training process. Experiments are conducted on Tiny-ImageNet at 120 seen classes with 50 labels for each class. We employ the following methods: (1) Labeled Only (w/o OS data), an SL method only trained with labeled data; (2) OpenMatch [18] (w/ all OS data), an OSSL method trained with all OS data; and (3) WiseOpen-L on top of OpenMatch(w/ selected OS data), an OSSL method trained with selected OS data.

- Certain OS data **(friendly data)** can **enhance** the ID classification accuracy.
- **Excluding** certain **unfriendly OS data** can further improve the ID classification performance, revealing that the **selection of OS training data is essential** for the OSSL task.

- **Wise Open-set Semi-supervised Learning(WiseOpen)**
- **Two practical and economic variants:WiseOpen-Economic (WiseOpen-E)/ WiseOpen-Loss (WiseOpen-L)**

Introduction

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

- Main contributions

- From the perspective of learning theory, we put forward an insight into the necessity of selectively leveraging the friendly open-set data in OSSL scenarios.
- We propose a robust general OSSL framework WiseOpen that employs GV-SM to wisely select friendly open-set data. This provides the OSSL community with a plug-and-play module to enhance the models' performance.
- We further provide WiseOpen-E and WiseOpen-L as two practical variants of WiseOpen, which can make the selection procedure more computation-friendly while still yielding performance improvements.
- The effectiveness of our proposed WiseOpen and its variants is demonstrated by extensive experiments on three popular benchmark datasets.

Related Works

(a) Semi-supervised Learning

- **MixMatch/FixMatch**
- Jointly employs consistency regularization and pseudo-labeling techniques and applies a fixed threshold for selecting high-confident unlabeled data to train the model.
- **Dash/FlexMatch/FreeMatch**
- Further explore how to determine the suitable confidence thresholds according to model's learning status so that better exploit unlabeled data for better performance.

(b) Open-set Semi-supervised Learning

- **OpenMatch**
- Adopts one-vs-all classifiers with soft open-set consistency regularization and incorporates FixMatch to handle the OSSL tasks.
- **IOMatch**
- Select confident pseudo-ID data while calculating unlabeled inlier loss, and exclude unlabeled data with low confidence in open-set predictions while calculating open-set loss where all unseen classes are regarded as one single class.

(c) Out-of-distribution Detection

- Existing OOD detection methods usually acquire **abundant labeled data** from seen classes and some methods even additionally **utilize external OOD knowledge** in the training phase.

Method

TABLE 1: Frequently used notations along with their mathematical meaning.

Notation	Mathematical Meaning
K	Number of the seen classes.
$\mathcal{S} = \{(\mathbf{x}_i^l, \mathbf{y}_i)\}_{i=1}^{N_l}$	Labeled training dataset containing N_l labeled pairs$(\mathbf{x}_i^l, \mathbf{y}_i)$.
$\mathcal{U} = \{\mathbf{x}_i^u\}_{i=1}^{N_u}$	Original unlabeled training dataset containing N_u instance $\mathbf{x}_i^u$.
$\mathcal{U}_t$	Selected unlabeled subset in the t-th epoch.
$\mathcal{L}, \mathcal{L}_s, \mathcal{L}_u$	The overall loss, supervised loss, and unsupervised loss.
θ	Parameters of the model.
$g(\theta)$	Stochastic gradient of loss function computed at θ.
E[·]	Mathematical expectation of some random variable.
ERB(·)	The excess risk bound given the model parameters.
$\lvert\cdot\rvert$	Cardinality of the given set.

overall objective function $\mathcal{L}$ of OSSL, no matter what specific techniques are applied, can be written as

$$\mathcal{L} = \mathcal{L}_s(\theta; \mathcal{S}) + \mathcal{L}_u(\theta; \mathcal{U}), \tag{1}$$

respectively. Then to train the model upon labeled training data set $\mathcal{S}$, OpenMatch will compute the following losses:

$$\mathcal{L}_{ce}(\theta; \mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{(\mathbf{x}_i^l, \mathbf{y}_i) \in \mathcal{S}} H(\mathbf{y}_i, \mathbf{p}(\mathbf{x}_i^l)), \tag{2}$$

$$\mathcal{L}_{ova}(\theta; \mathcal{S}) = -\frac{1}{|\mathcal{S}|} \sum_{(\mathbf{x}_i^l, \mathbf{y}_i) \in \mathcal{S}} \log q_0^{y_i}(\mathbf{x}_i^l) + \min_{k \neq y_i} \log q_1^k(\mathbf{x}_i^l), \tag{3}$$

$\mathbf{p}(\theta; \mathbf{x}_i^l)$: by traditional ID classifier

$\mathbf{q}^k(\mathbf{x}_i^l) = (q_0^k(\mathbf{x}_i^l), q_1^k(\mathbf{x}_i^l))$: by OOD detector

Method

$$\mathcal{L}_{em}(\theta;\mathcal{U}) = -\frac{1}{|\mathcal{U}|}\sum_{\mathbf{x}_i^u\in\mathcal{U}}\sum_{j=0}^{1}\sum_{k=1}^{K}\mathbf{q}^k(\alpha_j(\mathbf{x}_i^u))\log\mathbf{q}^k(\alpha_j(\mathbf{x}_i^u)), \tag{4}$$

$$\mathcal{L}_{oc}(\theta;\mathcal{U}) = \frac{1}{|\mathcal{U}|}\sum_{\mathbf{x}_i^u\in\mathcal{U}}\sum_{k=1}^{K}\|\mathbf{q}^k(\alpha_0(\mathbf{x}_i^u))-\mathbf{q}^k(\alpha_1(\mathbf{x}_i^u))\|_2^2. \tag{5}$$

Moreover, it adopts FixMatch over the pseudo-ID instances which can be formulated as

$$\mathcal{L}_{fm}(\theta;\mathcal{U}) = \frac{1}{|\mathcal{U}|}\sum_{\mathbf{x}_i^u\in\mathcal{U}}\mathcal{M}(\alpha(\mathbf{x}_i^u))H(\hat{\mathbf{y}}_i^u,\mathbf{p}(\mathcal{A}(\mathbf{x}_i^u))), \tag{6}$$

objective function of OpenMatch can be written as

$$\mathcal{L} = \underbrace{\mathcal{L}_{ce}(\theta;\mathcal{S}) + \mathcal{L}_{ova}(\theta;\mathcal{S})}_{\mathcal{L}_s(\theta;\mathcal{S})} + \underbrace{\lambda_1\mathcal{L}_{em}(\theta;\mathcal{U}) + \lambda_2\mathcal{L}_{oc}(\theta;\mathcal{U}) + \lambda_3\mathcal{L}_{fm}(\theta;\mathcal{U})}_{\mathcal{L}_u(\theta;\mathcal{U})}, \tag{7}$$

$$\mathcal{M}(\alpha(\mathbf{x}_i^u)) = \mathbb{I}(q_0^{\hat{y}_i^u}(\alpha(\mathbf{x}_i^u)) > 0.5)\cdot\mathbb{I}(\max(\mathbf{p}(\alpha(\mathbf{x}_i^u))) > \rho)$$

$$\hat{y}_i^u = \arg\min\mathbf{p}(\alpha(\mathbf{x}_i^u))$$

Method

risk minimization (RM) problem

$$\min_{\theta} \mathcal{L}(\theta) := \mathrm{E}_{\zeta \sim \mathcal{D}}[\ell(\theta; \zeta)], \quad (8)$$

unknown distribution D

empirical risk minimization (ERM) problem

$$\min_{\theta} \widehat{\mathcal{L}}(\theta) := \frac{1}{|\widehat{\mathcal{D}}|} \sum_{\zeta \in \widehat{\mathcal{D}}} \ell(\theta; \zeta)$$

SGD

$$\theta_{t+1} = \theta_t - \eta \widehat{g}(\theta_t)$$

ID and OOD data in $\widehat{\mathcal{D}}$

$$\widehat{g}(\theta_t) = \lambda g_{\mathrm{id}}(\theta_t) + (1-\lambda)(\tau g_{\mathrm{fr}}(\theta_t) + (1-\tau) g_{\mathrm{uf}}(\theta_t)), \quad (9)$$

use the gradient variances to measure the distance between labeled data and open-set data

Assumptions

Assumption 1 (Bounded variance [49]). *The stochastic gradient is unbiased,* $\mathrm{E}[\widehat{g}(\theta)] = \nabla\mathcal{L}(\theta)$. *The stochastic gradient* $g_{id}(\theta)$ *is variance bounded, i.e., there exists a constant* $\sigma^2 > 0$, *such that*

$$\mathrm{E}[\|g_{id}(\theta) - \nabla\mathcal{L}(\theta)\|^2] \leq \sigma^2.$$

Assumption 3 (Smoothness [52]). $\mathcal{L}(\theta)$ *is smooth with an L-Lipchitz continuous gradient, i.e., it is differentiable and there exists a constant* $L > 0$ *such that*

$$\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\| \leq L\|\theta - \theta'\|.$$

This is equivalent to

$$\mathcal{L}(\theta) - \mathcal{L}(\theta') \leq \langle\mathcal{L}(\theta'), \theta - \theta'\rangle + \frac{L}{2}\|\theta - \theta'\|^2.$$

Assumption 4 (Polyak-Łojasiewicz condition [53]). *There exists a constant* $\mu > 0$ *such that*

$$2\mu(\mathcal{L}(\theta) - \mathcal{L}(\theta_*)) \leq \|\nabla\mathcal{L}(\theta)\|^2,$$

where $\theta_* \in \arg\min_\theta \mathcal{L}(\theta)$ *is a optimal solution.*

Assumption 2 (Weak Growth Condition [50], [51]). *The stochastic gradient of* $g_{fr}(\theta)$ *and* $g_{uf}(\theta)$ *are variance bounded, i.e., there exists a constant* $\sigma^2 > 0$, *such that*

$$\mathrm{E}[\|g_{fr}(\theta) - \nabla\mathcal{L}(\theta)\|^2] \leq \frac{\epsilon}{2}\|\nabla\mathcal{L}(\theta)\|^2 + \sigma^2.$$
$$\mathrm{E}[\|g_{uf}(\theta) - \nabla\mathcal{L}(\theta)\|^2] \leq \frac{\nu}{2}\|\nabla\mathcal{L}(\theta)\|^2 + \sigma^2.$$

where $\epsilon > 0$ *is a small constant and* $\nu \gg 1$ *is a large enough constant.*

Since $\nu \gg 1$ is large enough, we can consider that the variance for $g_{\mathrm{uf}}(\theta)$ is much larger than the variance for $g_{\mathrm{fr}}(\theta)$. We consider the friendly data to have small variance so we suppose $\epsilon > 0$ is small. Similarly, we consider $\nu \gg 1$ is large enough for unfriendly data. In generalization analysis, we are interested in the excess risk bound (ERB):

$$\mathrm{ERB}(\widehat{\theta}) := \mathcal{L}(\widehat{\theta}) - \mathcal{L}(\theta_*), \tag{10}$$

where $\widehat{\theta}$ is a solution obtained by an algorithm and $\theta_* \in \arg\min_\theta \mathcal{L}(\theta)$ is the optimal solution of problem (8). For the convenience of analysis, we make the following widely used assumptions for the loss function.

Method

Theorem 1. *Under assumptions 1, 2, 3, 4, we have the following ERB in expectation:*
(a) when all data are used: by setting $\eta \leq \frac{1}{(1-\lambda)(\tau\epsilon+(1-\tau)\nu)L}$, *we have*

$$ERB(\widehat{\theta}_{id+uf+fr}) \leq O\left(\mathcal{L}(\theta_0) - \mathcal{L}(\theta_*)\right);$$

(b) when only labeled data are used: by setting $\eta = \frac{2}{n\mu}\log\left(\frac{n\mu^2(\mathcal{L}(\theta_0)-\mathcal{L}(\theta_*))}{\sigma^2 L}\right)$, *we have*

$$\begin{aligned} ERB(\widehat{\theta}_{id}) &\leq \frac{L\sigma^2}{n\mu^2} + \frac{2L\sigma^2}{n\mu^2}\log\left(\frac{n\mu^2(\mathcal{L}(\theta_0) - \mathcal{L}(\theta_*))}{\sigma^2 L}\right) \\ &\leq O\left(\frac{\log(n)}{n}\right); \end{aligned}$$

(c) when labeled and friendly data are used: by setting $\eta = \frac{2}{(n+m)\mu}\log\left(\frac{(n+m)\mu^2(\mathcal{L}(\theta_0)-\mathcal{L}(\theta_*))}{\sigma^2 L}\right)$, *we have*

$$\begin{aligned} ERB(\widehat{\theta}_{id+fr}) &\leq \frac{L\sigma^2}{(n+m)\mu^2} \\ &\quad + \frac{2L\sigma^2}{(n+m)\mu^2}\log\left(\frac{(n+m)\mu^2(\mathcal{L}(\theta_0) - \mathcal{L}(\theta_*))}{\sigma^2 L}\right) \\ &\leq O\left(\frac{\log(n+m)}{(n+m)}\right), \end{aligned}$$

where n *and* m *are the sample sizes of labeled data and friendly data, respectively,* $\widetilde{O}(\cdot)$ *suppresses a logarithmic factor and constants.*

The results of Theorem 1 show that (1) the algorithm could not reduce the objective due to the large variance arising from unfriendly open-set data; (2) by using friendly open-set data, the algorithm could significantly reduce the objective, and it has better generalization by comparing with the one only using labeled data. The theoretical findings inspire us to design a selection method for wisely leveraging open-set data. Specifically, one may carefully select and use friendly open-set data during training progress to improve the performance of the learning task.

↓

Wisely leveraging open-set data

Method

- **Wise Selection Mechanism(WiseOpen)**

gradient variance-based selection mechanism(GV-SM): discard the unfriendly open-set data with large gradient variance

$$\mathcal{U}_t = \left\{ \mathbf{x}_i^u \in \mathcal{U} \mid \| g_{\mathbf{x}_i^u}(\theta_t) - \bar{g}(\theta_t) \| < \sqrt{\rho_t} \right\}, \qquad (11)$$

$\bar{g}(\theta_t)$: estimated expectation of the gradient of the overall objective function L

$$\bar{g}(\theta_t) = \frac{1}{N_l} \sum_{i=1}^{N_l} \frac{\partial \mathcal{L}_s(\theta_t; \mathbf{x}_i^l)}{\partial \theta_t}. \qquad (12)$$

$$g_{\mathbf{x}_i^u}(\theta_t) = \frac{\partial \mathcal{L}_u(\theta_t; \mathbf{x}_i^u)}{\partial \theta_t}. \qquad (13)$$

- **Wise Selection Mechanism(WiseOpen)**

Obtain ρ_t

Top-k : utilizes the k-th largest gradient variance among $\{g_{\mathbf{x}_i^u}(\theta_t)\}_i^{N_u}$

Otsu thresholding : adaptively determine ρ_t by maximizing the variance of $g_{\mathbf{x}_i^u}(\theta_t)$ between the selected and discarded open-set data clusters

$$\mathcal{L} = \underbrace{\mathcal{L}_{ce}(\theta_t;\mathcal{S}) + \mathcal{L}_{ova}(\theta_t;\mathcal{S})}_{\mathcal{L}_s(\theta_t;\mathcal{S})} + \underbrace{\lambda_1\mathcal{L}_{em}(\theta_t;\mathcal{U}_t) + \lambda_2\mathcal{L}_{oc}(\theta_t;\mathcal{U}_t) + \lambda_3\mathcal{L}_{fm}(\theta_t;\mathcal{U}_t)}_{\mathcal{L}_u(\theta_t;\mathcal{U}_t)}. \tag{15}$$

Method

TABLE 3: Models' training time of original OpenMatch and our proposed frameworks on top of OpenMatch. Experiments are conducted using CIFAR-100 with 100 labels on a single NVIDIA GeForce RTX 4090.

Algorithm	Training Time
OpenMatch	10h 59m
w/ WiseOpen †	43h 01m
w/ WiseOpen ‡	43h 05m
w/ WiseOpen-E †	13h 40m
w/ WiseOpen-E ‡	13h 43m
w/ WiseOpen-L †	11h 36m
w/ WiseOpen-L ‡	11h 43m

computationally expensive to calculate gradient variance

Method

- **WiseOpen-E:simply sets an interval e_s of updating the selecting result of open-set data.**

$$\theta_t = \theta_{t-1} - \eta g(\theta_{t-1})$$

after m epochs

$$\theta_{t+m} = \theta_{t-1} - \eta \sum_{k=0}^{m} g(\theta_{t-1+k})$$

Assumption 3 (Smoothness [52]). *$\mathcal{L}(\theta)$ is smooth with an L-Lipchitz continuous gradient, i.e., it is differentiable and there exists a constant $L > 0$ such that*

$$\|\nabla\mathcal{L}(\theta) - \nabla\mathcal{L}(\theta')\| \leq L\|\theta - \theta'\|.$$

This is equivalent to

$$\mathcal{L}(\theta) - \mathcal{L}(\theta') \leq \langle\mathcal{L}(\theta'), \theta - \theta'\rangle + \frac{L}{2}\|\theta - \theta'\|^2.$$

By the condition of smoothness variance with parameter L', we have

$$\|g(\theta_{t+m}) - g(\theta_t)\| \leq L'\|\theta_{t+m} - \theta_t\| = \eta L' \left\|\sum_{k=1}^{m} g(\theta_{t-1+k})\right\|. \quad (16)$$

- **WiseOpen-L:applies a loss-based selection mechanism (L-SM) that selects the friendly open-set data with smaller loss values to construct U_t**

Assumption 4 (Polyak-Łojasiewicz condition [53]). *There exists a constant $\mu > 0$ such that*

$$2\mu(\mathcal{L}(\theta) - \mathcal{L}(\theta_*)) \leq \|\nabla \mathcal{L}(\theta)\|^2,$$

where $\theta_* \in \arg\min_\theta \mathcal{L}(\theta)$ is a optimal solution.

$$\mathcal{U}_t = \{\mathbf{x}_i^u \in \mathcal{U} \mid \mathcal{L}_u(\theta_t; \mathbf{x}_i^u) < \rho_t'\}, \tag{19}$$

L-SM

$$\ell(\theta) \leq \frac{1}{2\mu}\|\nabla \ell(\theta)\|^2 + \ell(\theta_*), \tag{17}$$

apply function $\mathcal{L}_u(\theta; \mathbf{x}_i^u)$.

$$\mathcal{L}_u(\theta_t; \mathbf{x}_i^u) \leq \frac{1}{2\mu}\|g_{\mathbf{x}_i^u}(\theta_t)\|^2 + \mathcal{L}_u(\theta_*; \mathbf{x}_i^u).$$

Once $\|g_{\mathbf{x}_i^u}(\theta_t) - \bar{g}(\theta_t)\| \leq \sqrt{\rho_t}$ holds in (11), then by $\|g_{\mathbf{x}_i^u}(\theta_t) - \bar{g}(\theta_t)\| \geq \|g_{\mathbf{x}_i^u}(\theta_t)\| - \|\bar{g}(\theta_t)\|$ we have $\|g_{\mathbf{x}_i^u}(\theta_t)\| \leq \sqrt{\rho_t} + \|\bar{g}(\theta_t)\|$, so that

$$\mathcal{L}_u(\theta_t; \mathbf{x}_i^u) \leq \frac{(\sqrt{\rho_t} + \|\bar{g}(\theta_t)\|)^2}{2\mu} + \mathcal{L}_u(\theta_*; \mathbf{x}_i^u). \tag{18}$$

Method

- **Pseudo-code**

Algorithm 1: WiseOpen Family.

Input: Labeled data $\mathcal{S}$, unlabeled data $\mathcal{U}$, model parameters θ, epoch E_{max}, iteration I_{max}, learning rate η, selection interval e_s.

```
for t ← 1 to E_max do
    if t % e_s == 0 then
        Obtain U_t according to Eq.11 or Eq.19;
    else
        Obtain U_t = U_{t-1};
    end
    for iter ← 1 to I_max do
        Sample batches B_l ∈ S and B_u ∈ U_t;
        Compute L ← L_s(θ; B_l) + L_u(θ; B_u);
        Update θ ← θ − η ∂L/∂θ;
    end
end
```

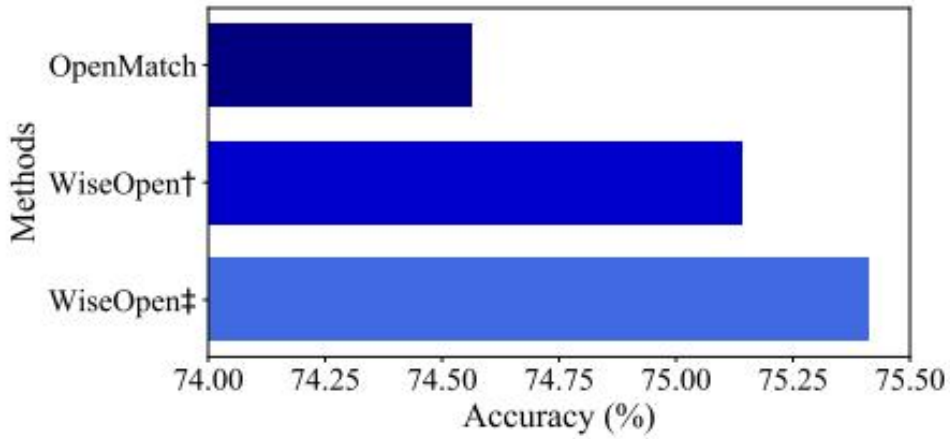
GV-SM:

$$\mathcal{U}_t = \left\{ \mathbf{x}_i^u \in \mathcal{U} \mid \| g_{\mathbf{x}_i^u}(\theta_t) - \bar{g}(\theta_t) \| < \sqrt{\rho_t} \right\}, \tag{11}$$

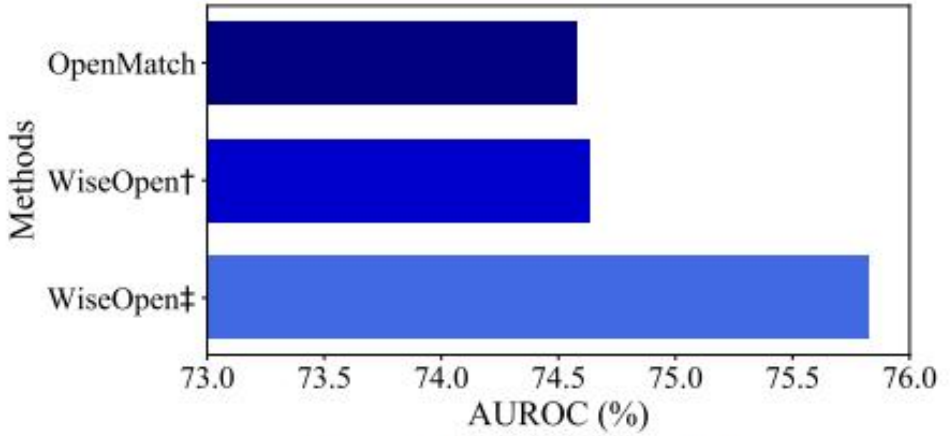
L-SM:

$$\mathcal{U}_t = \left\{ \mathbf{x}_i^u \in \mathcal{U} \mid \mathcal{L}_u(\theta_t; \mathbf{x}_i^u) < \rho_t' \right\}, \tag{19}$$

Experiments



(a) Evaluation of ID classification.



(b) Evaluation of OOD detection.

TABLE 3: Models' training time of original OpenMatch and our proposed frameworks on top of OpenMatch. Experiments are conducted using CIFAR-100 with 100 labels on a single NVIDIA GeForce RTX 4090.

Algorithm	Training Time
OpenMatch	10h 59m
w/ WiseOpen †	43h 01m
w/ WiseOpen ‡	43h 05m
w/ WiseOpen-E †	13h 40m
w/ WiseOpen-E ‡	13h 43m
w/ WiseOpen-L †	11h 36m
w/ WiseOpen-L ‡	11h 43m

Experiments

TABLE 4: Comparison of ID classification accuracy (in %, mean ± standard deviation) on CIFAR-10, CIFAR-100, and Tiny ImageNet with varying labels per seen class. † means using Top-k threshold while ‡ means using Otsu threshold. Each block consists of the results of the baseline with or without the variants of WiseOpen and the improvements that our proposed frameworks can make. The best results for each data setting in each block are in bold.

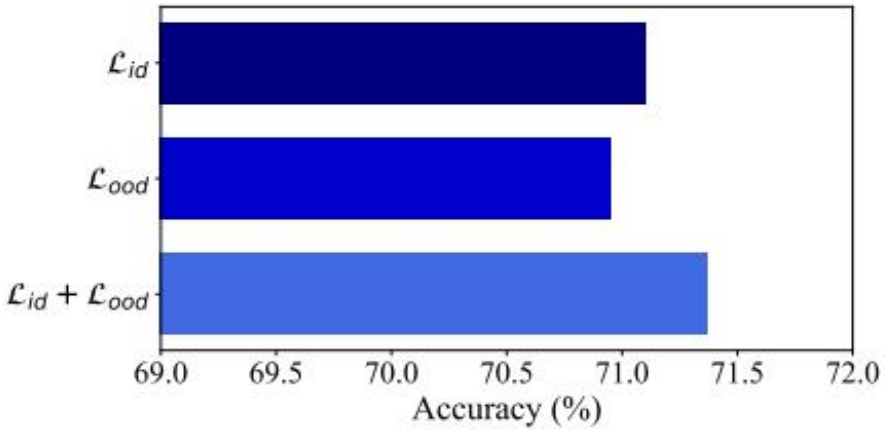
	CIFAR-10			CIFAR-100		Tiny-ImageNet
Algorithms	50 labels	100 labels	400 labels	50 labels	100 labels	50 labels
Labeled Only	63.16±0.87	67.26±1.08	83.67±0.29	60.15±0.20	64.96±0.29	42.31±0.43
FixMatch [14]	90.48±0.01	**92.61±0.16**	93.68±0.46	70.16±0.48	**74.17±0.22**	**45.11±0.53**
w/ WiseOpen-E †	91.33±0.15	**92.61±0.32**	93.67±0.20	**70.53±0.48**	74.13±0.41	44.90±0.63
w/ WiseOpen-E ‡	**91.52±0.44**	92.54±0.11	**93.80±0.31**	69.86±0.16	74.12±0.38	44.98±0.57
Δ (mean)	+0.94	-0.03	+0.06	+0.04	-0.04	-0.17
Δ (max)	+1.04	0.00	+0.12	+0.38	-0.04	-0.13
FreeMatch [27]	85.43±0.64	88.44±0.36	89.47±0.29	65.38±0.80	70.20±0.27	42.16±0.84
w/ WiseOpen-E †	**86.09±2.00**	88.36±0.80	89.76±0.75	**65.68±0.52**	**70.32±0.40**	**42.74±0.06**
w/ WiseOpen-E ‡	85.71±0.33	**88.71±0.36**	**90.37±0.76**	65.52±0.47	70.13±0.12	42.49±0.55
Δ (mean)	+0.47	+0.09	+0.60	+0.22	+0.02	+0.46
Δ (max)	+0.66	+0.27	+0.90	+0.30	+0.12	+0.58
MTC [26]	79.00±1.73	80.51±1.67	89.03±0.93	64.22±0.61	70.22±0.57	39.57±0.17
w/ WiseOpen-E †	81.37±2.71	82.58±1.51	**89.73±0.34**	**64.71±0.28**	70.33±0.12	**40.49±0.48**
w/ WiseOpen-E ‡	82.57±0.40	82.17±1.08	89.27±0.45	64.54±0.48	**70.42±0.16**	39.80±0.11
w/ WiseOpen-L †	81.34±1.89	83.45±1.45	89.23±0.87	64.68±0.83	70.16±0.76	39.83±0.28
w/ WiseOpen-L ‡	**82.65±0.32**	**85.69±1.55**	89.54±0.49	64.39±0.40	70.34±0.26	38.99±0.38
Δ (mean)	+2.98	+2.97	+0.42	+0.36	+0.09	+0.21
Δ (max)	+3.65	+5.19	+0.70	+0.49	+0.20	+0.92
OpenMatch [18]	82.45±2.31	91.23±0.94	92.80±0.45	70.23±0.30	74.56±0.46	47.33±0.81
w/ WiseOpen-E †	83.69±1.59	**91.86±0.45**	93.11±0.50	70.93±0.66	75.14±0.33	49.45±0.31
w/ WiseOpen-E ‡	83.35±1.95	91.47±0.53	**93.23±0.34**	**71.67±0.38**	74.55±0.19	49.14±0.33
w/ WiseOpen-L †	83.45±0.95	91.82±0.37	93.12±0.27	71.23±0.59	**75.38±0.58**	**49.75±0.69**
w/ WiseOpen-L ‡	**84.69±0.76**	91.34±0.69	92.93±0.06	71.12±0.31	75.09±0.43	48.74±0.08
Δ (mean)	+1.34	+0.39	+0.30	+1.01	+0.48	+1.94
Δ (max)	+2.24	+0.63	+0.43	+1.44	+0.82	+2.42
IOMatch [37]	91.54±0.32	92.09±0.36	93.46±0.17	69.83±0.59	73.87±0.25	47.86±0.24
w/ WiseOpen-E †	91.78±0.17	**92.25±0.62**	**93.59±0.07**	**70.49±0.28**	74.24±0.41	47.93±0.19
w/ WiseOpen-E ‡	91.77±0.08	92.16±0.18	93.36±0.16	69.97±0.55	74.12±0.12	47.99±0.33
w/ WiseOpen-L †	**91.90±0.16**	**92.25±0.20**	93.56±0.25	70.26±0.55	**74.36±0.45**	48.49±0.40
w/ WiseOpen-L ‡	91.16±0.29	92.02±0.20	93.35±0.08	70.47±0.37	74.33±0.05	**49.18±0.40**
Δ (mean)	+0.11	+0.08	+0.00	+0.47	+0.40	+0.54
Δ (max)	+0.36	+0.16	+0.13	+0.67	+0.49	+1.32

Experiments

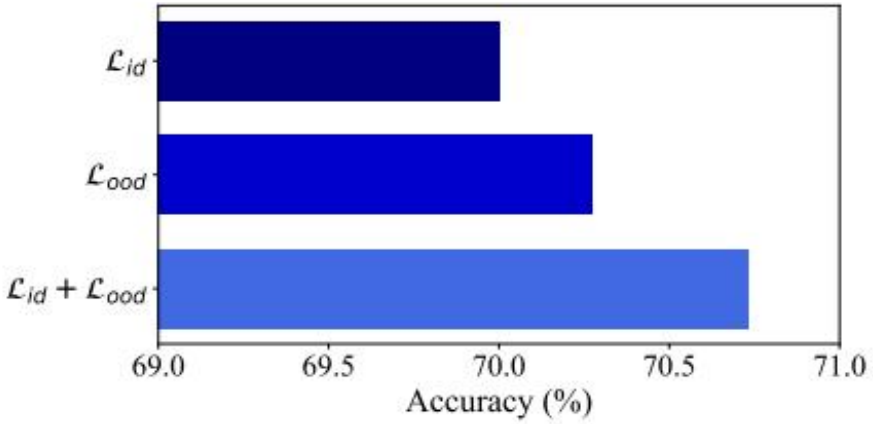
TABLE 5: Comparison of AUROC (in %, mean ± standard deviation) for evaluating OOD detection performance. Higher is better.

	CIFAR-10			CIFAR-100		Tiny-ImageNet
Algorithms	50 labels	100 labels	400 labels	50 labels	100 labels	50 labels
Labeled Only	56.15±1.81	59.58±2.25	69.95±0.60	67.86±0.71	70.04±0.52	61.49±0.38
FixMatch [14]	38.46±0.62	41.02±0.87	**48.49±1.41**	60.19±0.30	63.14±0.23	58.65±0.82
w/ WiseOpen-E †	**39.26±0.97**	41.20±1.56	47.53±0.44	**60.63±1.07**	62.57±0.12	**59.41±0.70**
w/ WiseOpen-E ‡	38.94±1.22	**42.26±1.10**	48.30±1.03	59.48±0.35	**63.28±0.52**	59.23±0.69
Δ (mean)	+0.64	+0.70	-0.58	-0.14	-0.22	+0.67
Δ (max)	+0.79	+1.23	-0.19	+0.44	+0.14	+0.76
FreeMatch [27]	45.94±1.84	**52.86±1.78**	**64.67±1.12**	**64.92±0.70**	68.71±0.19	**59.58±0.42**
w/ WiseOpen-E †	**47.38±2.03**	51.65±2.27	62.75±1.88	64.27±0.38	67.38±0.29	59.57±0.29
w/ WiseOpen-E ‡	46.46±2.36	52.79±2.37	63.94±2.67	64.22±0.89	**69.01±0.41**	58.84±0.86
Δ (mean)	+0.98	-0.64	-1.33	-0.68	-0.52	-0.38
Δ (max)	+1.44	-0.07	-0.73	-0.65	+0.30	-0.01
MTC [26]	77.77±1.10	80.36±2.13	87.02±0.91	65.40±0.60	64.58±0.26	60.71±0.55
w/ WiseOpen-E †	**79.02±0.35**	**82.02±1.98**	**88.67±0.76**	**66.78±1.79**	65.09±1.25	61.08±0.62
w/ WiseOpen-E ‡	78.10±0.36	79.08±1.74	86.59±2.45	65.97±0.91	64.41±1.25	61.29±0.08
w/ WiseOpen-L †	78.85±0.63	81.64±1.36	88.14±0.88	65.86±0.49	63.23±0.30	61.44±0.28
w/ WiseOpen-L ‡	78.45±0.39	81.65±0.70	86.10±1.71	65.41±1.42	**65.57±1.14**	**62.08±0.50**
Δ (mean)	+0.83	+0.73	+0.36	+0.60	-0.00	+0.76
Δ (max)	+1.25	+1.66	+1.65	+1.38	+0.99	+1.37
OpenMatch [18]	58.70±8.71	**55.60±5.09**	47.90±2.64	73.82±0.16	74.58±0.59	65.89±0.20
w/ WiseOpen-E †	**65.25±9.16**	49.97±5.71	**53.10±4.32**	75.23±0.49	76.12±0.72	66.41±0.15
w/ WiseOpen-E ‡	60.28±4.63	51.05±4.18	48.65±7.29	**75.24±0.34**	75.43±1.39	66.84±0.35
w/ WiseOpen-L †	55.33±4.99	49.70±4.45	44.14±3.55	74.59±0.40	**76.26±1.02**	66.71±0.57
w/ WiseOpen-L ‡	58.40±4.77	47.31±3.68	43.95±1.74	74.88±0.72	74.92±1.63	**67.33±0.21**
Δ (mean)	+1.11	-6.09	-0.44	+1.17	+1.10	+0.93
Δ (max)	+6.55	-4.55	+5.20	+1.42	+1.68	+1.44
IOMatch [37]	**44.54±0.29**	48.02±0.87	61.80±2.33	**67.44±0.88**	69.49±0.32	62.73±0.28
w/ WiseOpen-E †	42.51±2.22	48.34±0.96	61.49±2.17	65.99±0.14	69.36±0.48	**62.96±0.56**
w/ WiseOpen-E ‡	43.10±1.12	48.23±0.45	**63.45±1.04**	66.74±0.50	**69.67±0.10**	62.43±0.22
w/ WiseOpen-L †	44.23±1.74	**48.45±1.05**	60.58±1.18	67.04±0.43	69.43±0.53	62.66±0.60
w/ WiseOpen-L ‡	42.67±0.54	47.60±1.23	60.90±1.35	66.83±0.45	69.63±0.32	62.20±0.64
Δ (mean)	-1.41	+0.13	-0.20	-0.79	+0.03	-0.17
Δ (max)	-0.31	+0.43	+1.65	-0.40	+0.18	+0.23

Experiments



(a) WiseOpen-E ‡ on top of OpenMatch



(b) WiseOpen-E ‡ on top of IOMatch

TABLE 6: ID classification accuracy (in %) of WiseOpen-E and WiseOpen-L on top of OpenMatch and IOMatch in varying mismatching scenarios of CIFAR100 with 50 labels.

Mismatching ratios	0.4	0.5	0.6	0.7
OpenMatch	70.53	73.18	76.18	79.20
w/ WiseOpen-E †	71.85	73.68	77.40	78.93
w/ WiseOpen-E ‡	71.37	**74.06**	76.17	79.20
w/ WiseOpen-L †	**71.88**	73.76	**77.70**	**79.43**
w/ WiseOpen-L ‡	71.53	73.14	76.70	78.67
Δ (mean)	+1.13	+0.48	+0.81	-0.14
Δ (max)	+1.35	+0.88	+1.52	+0.23
IOMatch	70.53	72.74	74.98	77.63
w/ WiseOpen-E †	70.88	**73.20**	**75.52**	76.90
w/ WiseOpen-E ‡	70.73	73.04	75.35	77.20
w/ WiseOpen-L †	**71.02**	72.90	75.32	**78.27**
w/ WiseOpen-L ‡	70.93	73.02	75.07	78.13
Δ (mean)	+0.36	+0.30	+0.34	-0.00
Δ (max)	+0.49	+0.46	+0.54	+0.64

Experiments

TABLE 7: ID classification accuracy (in %) employing Adam optimizer and RMSProp optimizer. 50 labels per seen class are utilized in training models.

	CIFAR-10		CIFAR-100		Tiny-ImageNet	
Algorithms	Adam	RMSProp	Adam	RMSProp	Adam	RMSProp
OpenMatch	78.78	78.57	70.07	68.83	48.10	47.42
w/ WiseOpen-E †	73.10	76.88	69.77	69.52	47.98	**47.98**
w/ WiseOpen-E ‡	74.00	76.20	70.28	68.45	47.77	46.90
w/ WiseOpen-L †	**81.67**	78.93	71.12	68.40	48.18	47.13
w/ WiseOpen-L ‡	81.07	**82.40**	**71.47**	**69.97**	**48.35**	47.57
Δ (mean)	-1.32	+0.04	+0.59	+0.25	-0.03	-0.02
Δ (max)	+2.88	+3.83	+1.40	+1.13	+0.25	+0.57
IOMatch	90.27	91.78	70.60	70.32	45.35	45.28
w/ WiseOpen-E †	**91.37**	91.18	70.75	**70.73**	46.33	45.03
w/ WiseOpen-E ‡	90.55	91.83	**70.85**	70.53	46.13	44.63
w/ WiseOpen-L †	90.00	**92.32**	70.83	70.62	46.00	**45.30**
w/ WiseOpen-L ‡	89.75	91.05	70.68	70.70	**46.40**	45.23
Δ (mean)	+0.15	-0.19	+0.18	+0.33	+0.87	-0.23
Δ (max)	+1.10	+0.53	+0.25	+0.42	+1.05	+0.02

Experiments

TABLE 8: Evaluation of OOD detection on OOD data unseen in the training set (AUROC in %). Models are trained on Tiny-ImageNet with 50 labeled data per class.

	Unseen OOD Datasets						
Algorithms	LSUN	DTD	CUB	Flowers	Caltech	Dogs	MEAN
MTC	37.51±1.40	35.25±2.60	47.91±3.48	52.28±2.79	47.49±2.93	40.24±4.99	43.45±6.94
w/ WiseOpen-E †	39.39±7.39	37.73±2.75	48.70±0.72	55.24±3.04	51.69±0.79	44.01±3.36	46.13±7.36
w/ WiseOpen-E ‡	41.10±11.44	37.17±1.31	48.38±4.21	49.60±3.29	50.00±3.49	36.76±1.44	43.84±7.85
w/ WiseOpen-L †	34.82±2.45	35.93±0.86	50.06±3.25	54.41±6.50	50.87±2.83	43.24±1.65	44.89±8.25
w/ WiseOpen-L ‡	**45.43±15.31**	**47.81±2.07**	**58.03±4.75**	**60.51±2.96**	**57.26±4.41**	**48.85±1.24**	**52.98±9.06**
Δ (mean)	+2.68	+4.41	+3.38	+2.66	+4.96	+2.97	+3.51
Δ (max)	+7.92	+12.56	+10.11	+8.23	+9.77	+8.61	+9.53
OpenMatch	53.06±2.72	46.84±0.20	57.04±0.15	55.88±1.41	60.00±0.92	**61.13±0.67**	55.66±4.93
w/ WiseOpen-E †	54.38±1.24	47.68±1.38	58.45±1.14	55.72±2.30	61.84±1.06	59.77±1.34	56.31±4.81
w/ WiseOpen-E ‡	**56.41±0.98**	49.20±2.01	59.57±1.62	57.02±2.02	62.43±0.67	59.61±1.31	57.37±4.42
w/ WiseOpen-L †	56.22±0.78	**49.60±0.88**	59.39±0.20	59.62±1.17	**62.97±0.61**	59.81±1.46	**57.93±4.31**
w/ WiseOpen-L ‡	56.01±2.64	49.54±3.20	**59.97±0.14**	**59.81±1.91**	62.10±1.23	58.89±1.53	57.72±4.56
Δ (mean)	+2.70	+2.16	+2.30	+2.16	+2.33	-1.61	+1.68
Δ (max)	+3.35	+2.76	+2.93	+3.93	+2.97	-1.32	+2.28
IOMatch	63.88±1.76	56.53±2.11	63.10±2.07	64.52±4.22	63.71±1.15	**63.71±0.53**	62.57±3.56
w/ WiseOpen-E †	**67.28±0.59**	57.46±1.65	65.14±1.54	67.37±0.38	62.87±0.34	62.77±1.48	63.81±3.57
w/ WiseOpen-E ‡	65.47±1.67	**59.75±0.60**	64.23±1.03	66.29±3.21	63.31±0.64	62.98±1.39	63.67±2.68
w/ WiseOpen-L †	66.16±3.31	57.07±2.41	**66.56±0.84**	67.04±2.40	**63.73±1.03**	63.41±0.98	**63.99±3.96**
w/ WiseOpen-L ‡	64.93±2.04	57.93±2.35	64.18±1.56	**68.85±0.69**	63.65±0.42	62.93±1.09	63.75±3.56
Δ (mean)	+2.08	+1.52	+1.93	+2.86	-0.32	-0.69	+1.23
Δ (max)	+3.40	+3.22	+3.47	+4.33	+0.02	-0.30	+1.42

Thanks