Towards a Theoretical Understanding of Semi-Supervised Learning under Class Distribution Mismatch

Pan Du, Suyun Zhao, Puhui Tan, Zisen Sheng, Zeyu Gan, Hong Chen, Cuiping Li

***Abstract*— Semi-supervised learning (SSL) confronts a formidable challenge under class distribution mismatch, wherein unlabeled data contain numerous categories absent in the labeled dataset. Traditional SSL methods undergo performance deterioration in such mismatch scenarios due to the invasion of those instances from unknown categories. Despite some technical efforts to enhance SSL by mitigating the invasion, the profound theoretical analysis of SSL under class distribution mismatch is still under study. Accordingly, in this work, we propose Bi-Objective Optimization Mechanism (BOOM) to theoretically analyze the excess risk between the empirical optimal solution and the population-level optimal solution. Specifically, BOOM reveals that the SSL error is the essential contributor behind excess risk, resulting from both the pseudo-labeling error and invasion error. Meanwhile, BOOM unveils that the optimization objectives of SSL under mismatch are binary: high-quality pseudo-labels and adaptive weights on the unlabeled instances, which contribute to alleviating the pseudo-labeling error and the invasion error, respectively. Moreover, BOOM explicitly discovers the fundamental factors crucial for optimizing the bi-objectives, guided by which an approach is then proposed as a strong baseline for SSL under mismatch. Extensive experiments on benchmark and real datasets confirm the effectiveness of our proposed algorithm.**

***Index Terms*—Semi-supervised learning, class distribution mismatch, excess risk, contrastive learning.**

I. Introduction

DEEP neural networks (DNNs) have achieved remarkable success in fully-supervised learning tasks. However, sufficient labeled data are usually unavailable in real applications due to the expensive annotation cost or even domain-specific knowledge required [1], [2]. To overcome the scarcity of labeled data, semi-supervised learning (SSL) is studied by leveraging an abundance of unlabeled data on the fundamental assumption that both unlabeled and labeled data share the same label space [3]–[6]. Lots of theoretical analyses [2], [7]–[9] have claimed that unlabeled data can be beneficial for learning, provided that both labeled and unlabeled ones derive from the identical joint distribution, denoted as $P(X, Y)$. Note that X and Y are bounded random variables over input space $\mathcal{X}$ and label space $\mathcal{Y}$ respectively. Under such assumption, empirical evidence [4]–[6], [10] suggests that unlabeled data substantially enhance the performance of the target classifier.

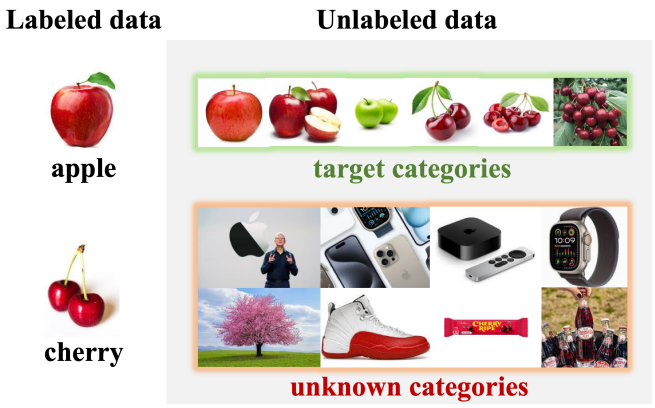


Fig. 1: Example of class distribution mismatch. The unlabeled data contains categories that are unseen in labeled ones.

Yet, a large number of SSL's real scenarios contradict the same-joint-distribution assumption, as the unlabeled data, collected in the open environment may contain numerous categories absent in the labeled data. For example, in the case of searching fruit images from the internet by keywords like "apple" and "cherry" (target categories), the collected images may include instances irrelevant to the target fruits, such as products associated with the Apple company (e.g., iPhone, Apple Watch), or products containing the word "cherry" (e.g., flowering cherry trees, Cheerwine Soft Drink). All these images are considered from unknown categories, as illustrated in Fig. 1. Similar occurrences are also observed in medical diagnoses [1] and the annotation of houses in remote-sensing images [11]. Obviously, the problems in such scenarios, commonly called class distribution mismatch, exhibit mismatched label space between labeled and unlabeled data. As a result, the joint distribution of unlabeled data diverges from that of the labeled data. Traditional SSL methods, unaware that the unlabeled data violates the same-joint-distribution assumption, proceed to learn the labeling function $P(Y|X;\boldsymbol{\theta})$ parameterized by $\boldsymbol{\theta}$, based on the biased joint distribution to mimic the latent input-label relationship between $\mathcal{X}$ and $\mathcal{Y}$, resulting in severe

Pan Du, Zisen Sheng, Zeyu Gan, Hong Chen, and Cuiping Li are affiliated with School of Information, Renmin University of China, Beijing, 100872, China, and Engineering Research Center of Database and Business Intelligence, MOE, China.

Suyun Zhao is affiliated with Engineering Research Center of Database and Business Intelligence, MOE, China, and the Key Laboratory of Data Engineering and Knowledge Engineering, MOE, China (corresponding author, email: zhaosuyun@ruc.edu.cn).

Puhui Tan is affiliated with School of Statistics, Renmin University of China, Beijing, 100872, China.

The profile for this paper is provided in BOOM-Profile.

Access our research codebase at BOOM-Code.

Manuscript received February 08, 2024; revised November 27, 2024.

performance deterioration and performing even worse than when using only labeled data [12], [13].

Recent efforts on SSL under mismatch have been directed towards aligning the joint distribution between labeled and unlabeled data by refining the data under class distribution mismatch into matched settings [12]–[17]. However, the adjusted joint distribution is susceptible due to its heavy reliance on the detection mechanism for unknown categories. Once a set number of instances with unknown categories infiltrate the training process, the learned labeling function $P(Y|X;\boldsymbol{\theta})$ becomes unreliable, and the performance of SSL declines sharply. Empirically, some methods [12], [13], [15], [16], which detect those unknown categories by leveraging the target classifier, already exhibit comparable or even worse performance compared to the ones trained with limited labeled data. To mitigate such agnostic risks, it's crucial to conduct a theoretical analysis that systematically evaluates the bias between the learned labeling function $P(Y|X;\boldsymbol{\theta})$ under class distribution mismatch and the optimal labeling function $P(Y|X;\boldsymbol{\theta}^*)$ that parameterized by $\boldsymbol{\theta}^*$. Such analysis should also be beneficial in discovering the fundamental factors crucial to align the joint distribution between labeled and unlabeled data.

To bridge the theoretical gap of SSL under mismatch, we explore the excess risk between the empirical optimal solution on the polluted data and the population-level optimal solution on the population data [18]. It is impossible to substitute the empirical solution into the expected error formula because the two solutions exist in distinct data spaces. Accordingly, we propose a **B**i-**O**bjective **O**ptimization **M**echanism (BOOM), which decouples the SSL error, the predominant component of excess risk, into dual objectives and then unveils the fundamental factors crucial for tightening its upper bound. BOOM mainly comprises three theorems. First and most important, BOOM reveals that the essence of excess risk [19] under mismatch lies in the SSL error. Then, as depicted in Theorem 1, the SSL error is decoupled into the pseudo-labeling error and the invasion error. To alleviate these two errors, BOOM advocates the dual optimization objectives by Theorem 2: enhancing pseudo-label quality as well as adaptively assigning weights to unlabeled instances. Last but not least, Theorem 3 in BOOM unveils the fundamental factors to optimize these objectives, providing some valuable insights for algorithmic design. Guided by the theoretical results of BOOM, we propose an SSL method to tighten the SSL error bound, thereby establishing a strong baseline of SSL under mismatch.

The main contributions can be summarized as follows.

i) We propose the Bi-Objective Optimization Mechanism (BOOM), which analyzes excess risk from empirical, sampling bias-free, and prior-free perspectives. To our knowledge, this is the first comprehensive theoretical framework specifically tailored to SSL under class distribution mismatch.
ii) Guided by BOOM, we propose an SSL method under mismatch to optimize both annotation and weights on the unlabeled data. This method may serve as a strong baseline supported by theoretical analysis under class distribution mismatch.
iii) Extensive experiments on four benchmark datasets and one real dataset validate the proposed theoretical framework, BOOM, and demonstrate the effectiveness of our method.

This study is a follow-up to the ICCV2023 publication [20], providing three notable advancements compared to its predecessor. i) Compared to the conference version, where only the empirical perspective of population risk is concerned, this study develops a comprehensive theoretical framework named BOOM, as presented in the supplementary Section III. BOOM systematically and comprehensively investigates the excess risk of SSL under class distribution mismatch from empirical, sampling bias-free, and prior-free perspectives. Specifically, it analyzes the version space in fully supervised, semi-supervised and mismatched semi-supervised manners and then establishes three distinct upper bounds for SSL error. This framework provides profound insights into potential agnostic risks and then offers a strict theoretical foundation for SSL algorithms under class distribution mismatch. ii) In light of the third theorem of BOOM, we reformat the proposed method WAD in the conference version, as detailed in Section IV. This enhancement confirms WAD's strong theoretical foundations, qualifying it as a solid baseline for SSL under class distribution mismatch. iii) Our experimental evaluation has been extended to encompass a broader range of mismatch proportions, from 0% to 100%, along with additional datasets (including Tiny-Imagenet and a realistic controlled noise dataset), further comparisons with the latest methods, and enhanced visualizations, as detailed in Section V. Besides, the whole paper—particularly the introduction and conclusion—has undergone a thorough revision aimed at a better understanding of BOOM.

The remainder of this paper is structured as follows. We begin by reviewing previous related works in Section II. Following this, Section III introduces the notations and presents the theoretical framework, BOOM. Additionally, a BOOM-guided method is provided in Section IV. Section V continues by exhibiting the extensive experimental results, while Section VI serves as the conclusion of the paper.

II. Related Work

This section reviews SSL methods in both class distribution match and mismatch scenarios, along with existing theoretical analyses. For insights on contrastive learning and a comprehensive overview of each SSL method's strengths and limitations, please refer to Appendix I.

A. Traditional SSL methods

Semi-supervised learning (SSL) aims to leverage both labeled and unlabeled data for model training. Traditional SSL strategies include entropy minimization, consistency regularization, and pseudo-labeling. Entropy minimization [4] incorporates unlabeled data in supervised learning by minimizing the entropy of the predictions for unlabeled instances. It encourages the model to make low-entropy predictions for unlabeled data, thereby preventing the decision boundary from passing through regions of high data density. Another SSL technique, consistency regularization [5], [6], [10], aims to ensure that the model's predictions remain stable and robust across different

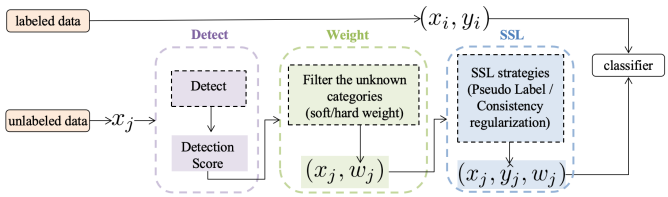


Fig. 2: The Detect-Weight-SSL paradigm of SSL methods under class distribution mismatch.

perturbations or views of the same instance. For example, the Π-Model [10], with the goal of minimizing the prediction discrepancy between the original instance and its perturbed counterpart, introduces stochastic perturbations to instances by employing data augmentation and dropout. As it depends on a single prediction from the target classifier, the Π-model suffers from instability. In contrast, Temporal Ensembling [5] concentrates on maintaining consistency between the Exponential Moving Average (EMA) of past predictions and current predictions. Furthermore, Virtual Adversarial Training (VAT) [21] generates adversarial perturbations that significantly alter the model's predictions and then minimizes the discrepancy between the perturbed and original predictions. As for the third SSL techniques, Pseudo-labeling-based approaches [22]–[25] mainly expand the labeled dataset by assigning pseudo labels to unlabeled instances. A pioneering method named Pseudo-Labeling [24] assigns high-confidence prediction labels as pseudo labels to unlabeled data. As a powerful SSL technique, FixMatch [25] first generates pseudo-labels based on the model's predictions on weakly augmented unlabeled images. If the pseudo-label is highly confident, it is aligned with the strongly augmented version of the same image using a cross-entropy loss function. We encourage interested readers to consult [26] for a detailed survey.

These conventional SSL methods perform well in match scenarios; however, they expose some intrinsic limitations when tackling unlabeled data with unknown categories.

B. SSL under Class Distribution Mismatch

To address the challenge of class distribution mismatch, most mismatched SSL methods [12]–[16], [27] adopt the detect-weight-SSL paradigm, as depicted in Fig. 2. This paradigm concentrates on the detection of instances associated with unknown categories, followed by filtering them through weights, thereby facilitating the SSL labeling techniques in handling mismatch issues.

Some methods [12], [13], [15], [16], following the detect-weight-SSL paradigm, leverage the target classifier to identify instances from unknown categories. For example, Deep Safe Semi-Supervised Learning (DS^{3}L) [12] leverages the target classifier to compute the prediction consistency loss between two augmented views of an unlabeled instance regarding instances with large discrepancies as unknown. Uncertainty Aware Self-Distillation (UASD) [13] filters out unknown categories by applying a confidence threshold to the averaged predictions of the temporally ensembled target classifier. With increased vulnerability, Class-aware Contrastive Semi-Supervised Learning (CCSSL) [15] and Simple but Strong Baseline (SSB) [16] directly apply a threshold to the maximum probability output of the target classifier to detect instances with unknown categories. However, a commendable aspect is that they further leverage instances with low weights to enhance the representation space in CCSSL [15] or to train an independent one-vs-all detector in SSB [16], improving the utility of the instances with unknown categories. After identifying the unknown categories, UASD [13] and SSB [16] generate pseudo-labels based on the output of the target classifier [28] to leverage instances with target categories, while DS^{3}L [12] and CCSSL [15] achieve this through consistency regularization strategies. Obviously, the detection performance is highly influenced by the classifier's effectiveness.

Unlike the aforementioned detection techniques, Trash to Treasure (T2T) [14] eliminates the dependency on the target classifier's output by introducing an additional cross-modal matching model. This model infers whether the embedding of the input image matches an assigned pseudo-label, thereby identifying instances associated with unknown categories. Subsequently, those large-weight instances are incorporated into training by leveraging consistency regularization. To improve T2T, OOD Semantic Pruning (OSP) [27] introduces a supplementary step to eliminate those pixels associated with unknown categories from target features. This is achieved by regularizing features from both target and unknown categories to be orthogonal. Although this approach further mitigates the negative effects of unknown categories at the pixel level, it may compromise the robustness of the learned model. Evidently, these SSL methods under mismatch are heavily reliant on the detection tasks, which may jeopardize rather than improve the performance of the target classifier. Should this detection fail, the learned target classifier breaks down. Therefore, it is crucial to uncover the underlying mechanism of SSL under mismatch to mitigate this agnostic risk.

Besides the detect-weight-SSL approaches, some methods [17], [29], [30] modify the fully connected layer of the target classifier, transforming the K-way classifier into a $K+1$-way model. In this setting, instances from unknown categories are treated as a distinct new class, with K representing the number of target categories. For example, a prototype network-based approach [17] adopts a distance-based function to identify instances associated with unknown categories, thereby generating new prototypes for these instances. Similarly, Simplifying Open-Set Semi-Supervised Learning with Joint Inliers and Outliers Utilization (IOMatch) [29] employs a group of one-vs-all classifiers to detect instances from unknown categories and in combination with the standard closed-set classifier to establish $K+1$-way classification targets accordingly. Additionally, the open-world method with uncertainty-based adaptive margin (ORCA) [30] assumes a known [31] or estimable [32] number of categories in the unlabeled data. It then learns a $K+M$-way target classifier using the proposed softmax function with an uncertainty-adaptive margin mechanism, where M denotes the number of unknown categories.

C. Theory of SSL

Theoretical advancements in SSL have primarily concentrated on investigating the potential role of unlabeled data [7]–

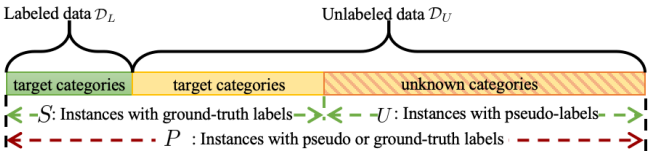


Fig. 3: The relations among data sets $\mathcal{D}_L$, $\mathcal{D}_U$, S, U, P. Note that $P \neq S \cup U$ due to the instances with target categories in $\mathcal{D}_U$ are assigned with pseudo-labels while not the ground-truth ones in S.

[9], [33]. From a critical perspective, Ben-David et al. [33] argue that SSL fails to guarantee the substantial benefits from unlabeled data without prior knowledge of label distribution. Therefore, researchers in the field of SSL consistently presume that either the cluster assumption or the manifold assumptions are intrinsic in unlabeled data. Under the clustering assumption, where the target function is thought locally smooth in the feature space [8], Rigollet et al. [7] establishes a tight upper bound on the generalization error in SSL classification. Moreover, under manifold assumption, where the target function is supposed on a low-dimensional manifold [8], Daniel Sanz-Alonso et al. [9] deduce that the unlabeled data are beneficial when utilizing graph-based methods in a Bayesian setting. However, these theoretical analyses lack sufficient guidance for practical algorithm design. As insightful results, Jia et al. [34] reveal that distribution discrepancies between labeled and unlabeled data stem from pseudo-label predictions and target predictions. To address this issue, one type of pseudo-labeling and weighting strategy is then proposed. Unfortunately, it just considers the covariate shift. Consequently, it is still promising to develop theoretical analysis in mismatch scenarios.

III. Bi-Objective Optimization Mechanism

To explore the class distribution mismatch problem in SSL, this section introduces the proposed theoretical framework, BOOM. It begins with the preliminaries and problem definition in Subsection III-A, followed by a comprehensive analysis of excess risk in Subsection III-B. Subsections III-C, III-D, and III-E further examine the theoretical results and practical guidelines for alleviating SSL errors. To make it easier to follow, Table I includes an abbreviated list of the main notations.

A. Preliminaries and Problem Defination

Preliminaries. Let ρ denote a population distribution defined over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ represent the input and output spaces, respectively, with $\mathcal{Y} \subseteq \mathbb{R}^K$. Suppose $\mathcal{H}$ is a hypothesis space, consisting of labeling functions $h : \mathcal{X} \to \mathcal{Y}$, i.e., parameterized classifiers, effectively mapping instances to labeling vectors in $\mathcal{Y}$, defined as $\mathcal{H} = \{h | \boldsymbol{y} = h_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{\theta} \in \Theta\}$, where $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{y} \in \mathcal{Y}$, and $\boldsymbol{\theta}$ represents the parameters governing the labeling function h, confined within the parameter space Θ. To measure the success of a prediction, we employ a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ to measure the dissimilarity between two elements within the output space $\mathcal{Y}$. Various loss functions satisfy the above definition, such as the cross-entropy and mean squared error (MSE). Given a noise-free dataset $T = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{|T|}$ that is drawn independently and identically distributed (i.i.d.) from a distribution ρ with a size of $|T|$, the objective of supervised learning is to derive a hypothesis $h \in \mathcal{H}$ that exhibits strong generalization performance to the ρ based on the dataset T, achieved by optimizing the empirical loss $\hat{\mathcal{L}}(h)$ as (1).

$$\hat{\mathcal{L}}(h) = \frac{1}{|T|} \sum_{i=1}^{|T|} \ell(h(\boldsymbol{x}_i), \boldsymbol{y}_i), \tag{1}$$

where the model that minimizes the empirical loss $\hat{\mathcal{L}}(h)$ is referred to as the empirical optimal solution.

Additionally, the corresponding expected loss is defined as (2).

$$\mathcal{L}(h) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\rho}[\ell(h(\boldsymbol{x}), \boldsymbol{y})], \tag{2}$$

where the model that minimizes the expected loss $\mathcal{L}(h)$ is referred to as the population-level optimal solution, denoted by h^*.

Then, the version space induced by T is defined as the subset of hypotheses space $\mathcal{H}$ where each hypothesis correctly predicts the labels of all instances in T. Formally, it is represented as $\{h \in \mathcal{H} \mid \forall(\boldsymbol{x}, \boldsymbol{y}) \in T, h(\boldsymbol{x}) = \boldsymbol{y}\}$ [35]. Since the hypotheses in the version space perfectly fit the training data, any hypothesis within this space is an empirical optimal solution.

Problem Definition. This study primarily investigates the problem of SSL under class distribution mismatch. Our ultimate goal is to develop a classifier for a K-class classification task that generalizes well to the instances from the distribution ρ. The training data comprises the limited number of labeled data $\mathcal{D}_L = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{m}$ that are i.i.d. sampled from distribution ρ with size of m and abundant unlabeled ones $\mathcal{D}_U = \{\boldsymbol{x}_i\}_{i=m+1}^{n}$ are i.i.d. drawn according to the marginal distribution ρ_X over $\mathcal{X}$ and an unknown marginal distribution ρ_{X_u} over $\mathcal{X}_u$ with size of $n - m$, typically $m \ll n - m$. Importantly, it should be noted that the input spaces $\mathcal{X}$ and $\mathcal{X}_u$ contain instances associated with distinct semantic categories.

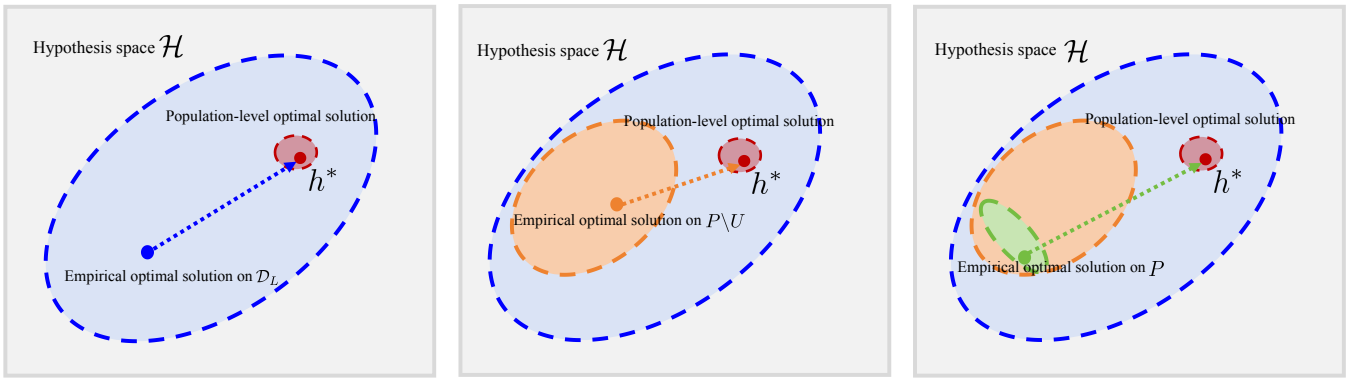
To leverage the unlabeled instances, this study adopts a pseudo-label-based SSL approach. Each instance in $\mathcal{D}_U$ is assigned a pseudo-label, thereby forming the dataset P with a size of n when combined with the labeled dataset $\mathcal{D}_L$. For convenience, we denote $S = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{s}$ as the instances in P, which are sampled from the marginal distribution ρ_X with their ground-truth labels, while U represents the set of instances in P, sampled from an unknown marginal distribution ρ_{X_u}, with their pseudo-labels. To explicitly delineate the relationships among P, S, U, $\mathcal{D}_L$, and $\mathcal{D}_U$, their relations are visually represented in Fig. 3. It is apparent that the dataset S is purely composed of instances from target categories. In contrast, the dataset P is potentially contaminated by incorrect pseudo-labels and invaded by instances from an unknown marginal distribution ρ_{X_u}.

B. Excess Risk Analysis

Excess risk refers to the gap between the population risk of a model trained on available data and that of the optimal model, which minimizes risk across the entire data distribution. To probe the essence of excess risk, we explore the relationships between the population-level optimal solution h^* on ρ and the empirical optimal solutions in supervised learning, SSL, and

TABLE I: List of notations.

Symbol	Definition	Symbol	Definition
$\mathcal{X}$	Input space	$\mathcal{X}_u$	Input space with distinct semantic categories from $\mathcal{X}$
$\mathcal{Y}$	Output space ($\subseteq \mathbb{R}^K$, K denotes the number of target categories.)	ρ	Population distribution over $\mathcal{X} \times \mathcal{Y}$
ρ_X	Marginal distribution over $\mathcal{X}$	ρ_{X_u}	Marginal distribution over $\mathcal{X}_u$
$\tilde{\rho}_X$	Empirical marginal distribution of instances sampled from $\mathcal{X}$	$\tilde{\rho}_{X_u}$	Empirical marginal distribution of instances sampled from $\mathcal{X}_u$
T	Dataset from ρ, size $\lvert T\rvert$	$\mathcal{D}_L$	Limited labeled data from ρ, size m
$\mathcal{D}_U$	Unlabeled data from $\rho_{\mathcal{X}}$ and ρ_{X_u}, size $n-m$	S	Target instances in $\mathcal{D}_L$ or $\mathcal{D}_U$ with ground-truth labels
P	Combined dataset from $\mathcal{D}_L$ and $\mathcal{D}_U$ with pseudo-labels, size n	U	Instances in P sampled from ρ_{X_u}
J	Set $\{(\boldsymbol{x}_i, \boldsymbol{y}_i, \hat{\boldsymbol{y}}_i)\}_{i=1}^{s}$, $(\boldsymbol{x}_i, \hat{\boldsymbol{y}}_i) \in S$, $\boldsymbol{x}_i \in \mathcal{X}$	h	Labeling function $h_{\boldsymbol{\theta}}$ with parameter $\boldsymbol{\theta}$
$\boldsymbol{\theta}$	Parameters of h in Θ	$\mathcal{H}$	Hypothesis space contains functions $h: \mathcal{X} \to \mathcal{Y}$
$\hat{h}$	Empirical optimal model on P using SSL methods for class distribution mismatch	h^*	Population-level optimal model on ρ
ℓ	Loss function $\ell: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$	λ^l	Lipschitz constant for ℓ
H	Bound for loss function ℓ	$\hat{\mathcal{L}}(h)$	Empirical loss under class distribution match
$\mathcal{L}(h)$	Expected loss under class distribution match	$\boldsymbol{w}$	Weight vector $\boldsymbol{w} = [w_1, ..., w_{\lvert P\rvert}]$
ϵ / $\hat{\epsilon}$ / κ	Fixed weight for labeled instances; $\hat{\epsilon} = \frac{n}{s}\epsilon$; $\kappa = 2$	$\tilde{\boldsymbol{w}}$	Fixed weight vector $\tilde{\boldsymbol{w}} = [\epsilon, ..., \epsilon]$
$\hat{\mathcal{L}}(h; \boldsymbol{w})$	Empirical loss under class distribution mismatch	$\mathcal{L}(h; \tilde{\boldsymbol{w}})$	Expected loss under class distribution mismatch
ω_t	Weight distribution for instances from ρ_X	ω_u	Weight distribution for instances from ρ_{X_u}
μ_t / σ_t^2 / τ_t	Expectation / Variance / Bound of weights from ω_t	μ_u / σ_u^2 / τ_u	Expectation / Variance / Bound of weights from ω_u
$\Phi(\boldsymbol{x})$	Labeling function	λ^u	Lipschitz constant for $\Phi(x)$
$\varphi(\boldsymbol{y})$	$\arg\max_k \boldsymbol{y}$	$\eta_k(\boldsymbol{x})$	Class-specific function: $p(\varphi(\boldsymbol{y}) = k \lvert \boldsymbol{x})$
λ^η	Lipschitz constant for $\eta_k(\boldsymbol{x})$	D	Distance bound between labeled $\boldsymbol{x}_j$ and unlabeled $\boldsymbol{x}_i$
L	Distance bound between $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$	R	Bound for $\lVert\omega_i - \epsilon\rVert_2$
$\Lambda(\boldsymbol{x})$	$\mathcal{X} \cup \mathcal{X}_u \to \mathbb{R}$	λ^w	Lipschitz constant for $\Lambda(\boldsymbol{x})$
$\hat{D}$	Distance bound between unlabeled $\boldsymbol{x}_i$ and labeled $\boldsymbol{x}_k$	$\hat{D} - \xi$	Distance bound between unlabeled $\boldsymbol{x}_i$ and labeled $\boldsymbol{x}_j^{'}$, $\boldsymbol{y}_k \neq \boldsymbol{y}_j^{'}$, $\xi \geq 0$
$\boldsymbol{x}_{i,st}$	Closest labeled instance to unlabeled $\boldsymbol{x}_i$	$\boldsymbol{x}_{i,nd}$	Closest labeled instance to unlabeled $\boldsymbol{x}_i$ where $\boldsymbol{y}_{i,nd} \neq \boldsymbol{y}_{i,st}$
$D_{i,st}$	Distance from unlabeled $\boldsymbol{x}_i$ to nearest labeled neighbor $\boldsymbol{x}_{i,st}$	$D_{i,nd}$	Distance from $\boldsymbol{x}_i$ to $\boldsymbol{x}_{i,nd}$
ξ	Width of the annulus centered at $\boldsymbol{x}_i$ with radii $\hat{D}$ and $\hat{D} - \xi$	$\xi_{i,st}$	Gap between $\hat{D}$ and $D_{i,st}$
$\boldsymbol{z}_{i,u}$	Feature of unlabeled instance	$\boldsymbol{z}_{i,st}^k$	Feature of $\boldsymbol{x}_{i,st}$ with ground-truth label k



(a) Supervised version space (blue region). (b) SSL version space (orange region). (c) Mismatched SSL version space (green region).

Fig. 4: Illustration of the relationship between the population-level optimal solution h^* on ρ and the empirical optimal solutions in supervised learning, SSL, and SSL under class distribution mismatch, respectively.

SSL under class distribution mismatch, as depicted in Fig. 4, respectively.

In supervised learning, the hypotheses that minimize the loss function (1) on the target distribution ρ and the labeled available dataset $\mathcal{D}_L$ form the population version space and supervised version space, respectively, as depicted in the red and blue regions in Fig. 4 (a). Any hypothesis contained in the population version space corresponds to a population-level optimal solution, while the one in the supervised version space corresponds to the empirical optimal solution. As the scale of labeled noise-free data $\mathcal{D}_L$ increases, the supervised version space shrinks and eventually converges to the population version space, i.e., the blue region shrinks to the red one. Hence, the empirical optimal solution may approach the population-level optimal solution in the case of abundant noise-free labeled data.

In traditional SSL, the empirical optimal solution is trained on labeled data and pseudo-labeled instances, i.e., $P\backslash U$. In such cases, the version space shrinks to the SSL version space on $P\backslash U$, depicted as the orange region Fig. 4 (b). Once errors occur in pseudo-labels, the orange region will not contain the red region. This shows that, although SSL may reduce the version space, it is susceptible to erroneous pseudo-labels, preventing it from converging to the population-level optimal solution.

In the mismatched scenario, incorporating unknown instances into the training process further shrinks the version space to the mismatched SSL version space, depicted as the green region in Fig. 4 (c). However, due to the invasion of these unknown instances, the green region heavily drifts from the red region, indicating a significant deviation of the mismatched SSL version space from the population version space. Therefore, under class distribution mismatch, leveraging uniform weights such as $\frac{1}{|T|}$, as shown in (1), may cause the learned target classifier to deviate from the optimal model. This assertion is supported by the theoretical analysis in i) and iv) of Corollary 1, as well as by the experiments presented in Subsection V-C.

To tackle the challenge of class distribution mismatch, it is imperative to calibrate the mismatched SSL version space to the population version space, thereby aligning the empirical optimal solution with the population-level optimal solution h^*. Such calibration will significantly enhance the model's generalization capability. Therefore, we optimize the weights $\boldsymbol{w}$ together with $\boldsymbol{\theta}$ to get an empirical optimal solution in mismatched SSL scenarios, and the general objective is defined

as:

$$\min_{\boldsymbol{w}\in W}\min_{h\in\mathcal{H}}\hat{\mathcal{L}}(h;\boldsymbol{w})=\frac{1}{|P|}\sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i)\in P}w_i\ell(h(\boldsymbol{x}_i),\boldsymbol{y}_i), \tag{3}$$

where $\boldsymbol{w}=\left[w_1,\cdots,w_{|P|}\right]$ is the vector of weights, and w_i remains constant at ϵ for the labeled instances; W is the parameter space for $\boldsymbol{w}$. In our framework, we relax the constraint $\sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i)\in P}w_i/|P|=1$ on $\boldsymbol{w}$, thereby expanding the feasibility range. Each element w_i can take any value under the assumption that the weights of instances from ρ_X and ρ_{X_u} follow the distribution ω_t with expectation μ_t and variance σ_t^2, and the distribution ω_u with expectation μ_u and variance σ_u^2, and restricted to τ_t and τ_u, respectively.

Then, the empirical optimal model $\hat{h}$, which is trained on the dataset P using SSL methods specifically designed for class distribution mismatch, and the population-level optimal model h^*, on ρ, are formulated in (4) and (5), respectively. Here, $\tilde{\boldsymbol{w}}$ represents a vector of weights where each element is a constant ϵ.

$$\hat{h}=\underset{\boldsymbol{w}\in W,h\in\mathcal{H},(\boldsymbol{x},\boldsymbol{y})\in P}{\arg\min}\hat{\mathcal{L}}(h;\boldsymbol{w}), \tag{4}$$

$$h^*=\underset{h\in\mathcal{H},(\boldsymbol{x},\boldsymbol{y})\sim\rho}{\arg\min}\mathcal{L}(h;\tilde{\boldsymbol{w}}), \tag{5}$$

which is equivalent to $h^*=\arg\min_{h\in\mathcal{H}}\mathcal{L}(h)$, as shown in (2). The population risk of a model is defined as:

$$\mathcal{L}(h;\tilde{\boldsymbol{w}})=\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\rho}[\epsilon\ell(h(\boldsymbol{x}),\boldsymbol{y})].$$

Subsequently, the excess risk between the empirical optimal solution on a polluted data P and the population-level optimal solution on pure population data is defined as $\mathcal{L}(\hat{h};\tilde{\boldsymbol{w}})-\mathcal{L}(h^*;\tilde{\boldsymbol{w}})$. A challenge arises because these solutions exist in distinct data spaces, rendering it impossible to substitute the empirical solution into the expected error formula. To overcome this challenge, we decompose the excess risk into three distinct components: the generalization gap of $\hat{h}$, the concentration error, and the SSL error, as shown in (6). For an in-depth derivation process, please refer to the Appendix IV-A.

$$\begin{aligned}&\mathcal{L}(\hat{h};\tilde{\boldsymbol{w}})-\mathcal{L}(h^*;\tilde{\boldsymbol{w}})\\&\leq\underbrace{\left|\mathcal{L}(\hat{h};\tilde{\boldsymbol{w}})-\hat{\mathcal{L}}_P(\hat{h};\boldsymbol{w})\right|}_{\textbf{Generalization gap}}+\underbrace{\left|\hat{\mathcal{L}}_S(h^*;\tilde{\boldsymbol{w}})-\mathcal{L}(h^*;\tilde{\boldsymbol{w}})\right|}_{\textbf{Concentration error}}\\&+\underbrace{\left|\hat{\mathcal{L}}_P(\hat{h};\boldsymbol{w})-\hat{\mathcal{L}}_S(h^*;\tilde{\boldsymbol{w}})\right|}_{\textbf{SSL error}},\end{aligned} \tag{6}$$

where $\hat{\mathcal{L}}_P$ and $\hat{\mathcal{L}}_S$ indicate the empirical loss on the datasets P and S, respectively.

Generalization gap. The generalization gap refers to the discrepancy between a learning model's performance on the available training data (P) and its performance on the population data drawn from the target distribution ρ. A smaller generalization gap indicates better generalization. Numerous theoretical studies have shown that the generalization gap of DNNs, which are employed in our work, can be effectively bounded [36], [37].

Concentration error. The concentration error indicates the deviation of the random variable $\hat{\mathcal{L}}_S(h^*;\tilde{\boldsymbol{w}})$ from its expectation $\mathcal{L}(h^*;\tilde{\boldsymbol{w}})$ [38]. As the size of the set S approaches infinity, $\hat{\mathcal{L}}_S(h^*;\tilde{\boldsymbol{w}})$ naturally concentrates around its expected value, minimizing the concentration error. Considering that the size of the set S is independent of the learning optimization, we need not delve into concentration error here. For more details about the concentration error, please refer to the Appendix IV-B.

SSL error. Unlike the generalization gap, the SSL error highlights discrepancies in model's performance between the contaminated training dataset P and the labeled pure one S. Undoubtedly, SSL error plays a critical role in contributing to the excess risk in SSL under class distribution mismatch.

C. Empirical Upper Bound for SSL Error

This subsection analyzes the upper bound of the SSL error at the empirical level and decouples it into the pseudo-labeling error and the invasion error, as stated in Theorem 1.

Theorem 1. *(Empirical upper bound for SSL error). Let $U=\{(\boldsymbol{x},\boldsymbol{y})|(\boldsymbol{x},\boldsymbol{y})\in P,\boldsymbol{x}\in\mathcal{X}_u\}$ denote a set of instances with size $n-s$, where $\boldsymbol{x}_i$ is sampled from ρ_{X_u} and follows the empirical distribution $\tilde{\rho}_{X_u}$. Similarly, let $\hat{\epsilon}=\frac{n}{s}\epsilon$, where ϵ is a constant and J represent the set $\{(\boldsymbol{x}_i,\boldsymbol{y}_i,\hat{\boldsymbol{y}}_i)\}_{i=1}^s$ with $(\boldsymbol{x}_i,\hat{\boldsymbol{y}}_i)\in P\backslash U$, $(\boldsymbol{x}_i,\boldsymbol{y}_i)\in S$, and $\boldsymbol{x}_i\in\mathcal{X}$, i.e., $\boldsymbol{x}_i$ is sampled from ρ_X and follows the empirical distribution $\tilde{\rho}_X$. Assume the loss function $\ell(\cdot,\boldsymbol{y})$ is Lipschitz continuous with a constant λ^l for all $\boldsymbol{y}$, h^*, and $\hat{h}$, and bounded by H for all $\mathcal{Y}\times\mathcal{Y}$. Given these assumptions, we can set the constant κ to 2, leading to the following result:*

$$\begin{aligned}&\left|\hat{\mathcal{L}}_P(\hat{h};\boldsymbol{w})-\hat{\mathcal{L}}_S(h^*;\tilde{\boldsymbol{w}})\right|\\&\leq\underbrace{\frac{1}{s}\sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i,\hat{\boldsymbol{y}}_i)\in J}\left[w_i\kappa\lambda^l\|\hat{\boldsymbol{y}}_i-\boldsymbol{y}_i\|_2+\|(w_i-\hat{\epsilon})H\|_2\right]}_{\textit{Pseudo-labeling error}}\\&+\underbrace{\frac{\kappa}{n-s}\sum_{(\boldsymbol{x}_i,\hat{\boldsymbol{y}}_i)\in U}w_iH}_{\textit{Invasion error}}.\end{aligned} \tag{7}$$

Corollary 1. *Under the same conditions as established in Theorem 1, we deduce the following consequences:*

i) If the weight of each instance in J equals $\hat{\epsilon}$:

If $\hat{\boldsymbol{y}}_i=\boldsymbol{y}_i$, for all instances in J:

$$\frac{1}{s}\sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i,\hat{\boldsymbol{y}}_i)\in J}\left[w_i\kappa\lambda^l\|\hat{\boldsymbol{y}}_i-\boldsymbol{y}_i\|_2+\|(w_i-\hat{\epsilon})H\|_2\right]=0,$$

while if $\hat{\boldsymbol{y}}_i\neq\boldsymbol{y}_i$, for all instances in J:

$$\begin{aligned}&\frac{1}{s}\sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i,\hat{\boldsymbol{y}}_i)\in J}\left[w_i\kappa\lambda^l\|\hat{\boldsymbol{y}}_i-\boldsymbol{y}_i\|_2+\|(w_i-\hat{\epsilon})H\|_2\right]\\&=\frac{1}{s}\sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i,\hat{\boldsymbol{y}}_i)\in J}\hat{\epsilon}\kappa\lambda^l\|\hat{\boldsymbol{y}}_i-\boldsymbol{y}_i\|_2.\end{aligned}$$

ii) If the weight of each instance in J equals 0, then:

$$\frac{1}{s}\sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i,\hat{\boldsymbol{y}}_i)\in J}\left[w_i\kappa\lambda^l\|\hat{\boldsymbol{y}}_i-\boldsymbol{y}_i\|_2+\|(w_i-\hat{\epsilon})H\|_2\right]=\hat{\epsilon}H.$$

iii) If the weight of each instance in U equals 0, then:

$$\frac{\kappa}{n-s}\sum_{(\boldsymbol{x}_i,\hat{\boldsymbol{y}}_i)\in U}w_iH=0.$$

iv) If the weight of each instance in U equals $\hat{\epsilon}$, then:

$$\frac{\kappa}{n-s}\sum_{(\boldsymbol{x}_i,\hat{\boldsymbol{y}}_i)\in U}w_iH=\hat{\epsilon}\kappa H.$$

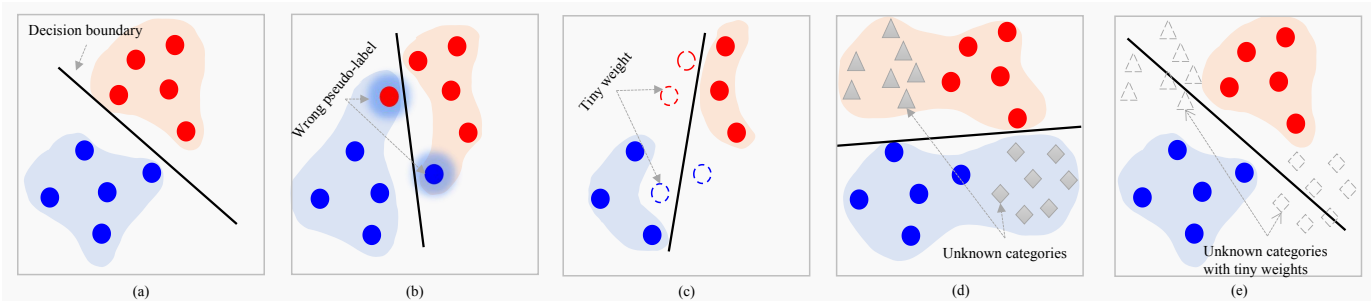


Fig. 5: The decision boundaries under various scenarios. ● and ● represent the instances with target categories, while ▲ and ◆ denote the instances with unknown categories. **(a)** illustrates the true decision boundary **(b)** represents the decision boundary in the case that the instances sampled from ρ_X are incorrectly annotated. **(c)** illustrates the decision boundary in the case that some instances from ρ_X are assigned with tiny weights. **(d)** displays the decision boundary if some instances with unknown categories invade. **(e)** depicts the decision boundary in the case that all instances with unknown categories are filtered out.

***Remark* 1.** Theorem 1 estimates the upper bound of the SSL error by leveraging the empirical distributions $\tilde{\rho}_X$ and $\tilde{\rho}_{X_u}$ of the instances sampled from the true distributions ρ_X and ρ_{X_u}, respectively. For a detailed proof of Theorem 1, please kindly refer to the Appendix IV-C.

i) Pseudo-labeling error. In Theorem 1, the pseudo-labeling error is defined on the empirical marginal distribution $\tilde{\rho}_X$. It quantifies the errors resulting from both incorrect pseudo-labeling and low participation in training. Concretely, it assesses the disparity between ground-truth labels and pseudo-labels by the term $\|\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i\|_2$. Meanwhile, the term $\|(w_i - \hat{\epsilon})H\|_2$ in SSL error penalizes the instances with lower weights. This is because the instances in J, from empirical distribution $\tilde{\rho}_X$, benefit to improve the target classifier. Provided that they are assigned too low weights to make them contribute less for training, a large penalty should be imposed to increase the pseudo-labeling error. Just as inferred from i) of Corollary 1, both incorrect pseudo-labels and those instances with low participation jointly contribute to the pseudo-labeling error. In addition, (b) and (c) of Fig. 5 illustrate that the decision boundaries are prone to fluctuate with the incorrect pseudo-labels as well as they are sensitive to the instances with tiny weights. Consequently, to alleviate the pseudo-labeling error, the pseudo-labels are supposed to align with their corresponding ground-truth labels. In the meanwhile, the weights of the unlabeled instances from the target empirical marginal distribution $\tilde{\rho}_X$ are supposed to approach $\hat{\epsilon} = \frac{n}{s}\epsilon$, to ensure their contributions for training.

ii) Invasion error. The invasion error arises in case that the instances from the empirical distribution $\tilde{\rho}_{X_u}$ infiltrate the training process of the target classifier. As illustrated in (d) of Fig. 5, this error may cause an erroneous decision boundary. According to Theorem 1 and iii) & iv) of Corollary 1, we reveal that the invasion error is controlled by the weights assigned to the instances from the empirical distribution $\tilde{\rho}_{X_u}$. Therefore, reducing the weights assigned to such instances with unknown categories may mitigate the invasion error, as depicted in (e) of Fig. 5.

iii) SSL error. The SSL error is twofold: the pseudo-labeling error and the invasion error. To effectively mitigate it under mismatch, there is a dual objective. One is to enhance the quality of pseudo-labels assigned to unlabeled instances from $\tilde{\rho}_X$, thereby alleviating the pseudo-labeling error. The other sub-objective concentrates on weight assignment to unlabeled instances. As shown in i) and iv) of Corollary 1, if instances from unknown categories are assigned a weight of zero, distinct from the weight $\hat{\epsilon}$ assigned to instances from target categories, then the upper bound of the SSL error decreases from $\frac{1}{s}\sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i,\hat{\boldsymbol{y}}_i)\in J}\hat{\epsilon}\kappa\lambda^l\|\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i\|_2 + \hat{\epsilon}\kappa H$ to $\frac{1}{s}\sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i,\hat{\boldsymbol{y}}_i)\in J}\hat{\epsilon}\kappa\lambda^l\|\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i\|_2$. Therefore, to mitigate the SSL error, we should assign tiny weights to instances from $\tilde{\rho}_{X_u}$ and large weights to instances from $\tilde{\rho}_X$ rather than uniform weights assigned to all unlabeled instances.

Summary of Theorem 1. Theorem 1 states the empirical upper bound of the SSL error established on the empirical distributions $\tilde{\rho}_X$ and $\tilde{\rho}_{X_u}$. It provides a fundamental theoretical analysis of SSL under class distribution mismatch. However, this empirical bound overlooks the sampling bias, where the empirical distribution of the sampled data deviates from the true sample distribution. Certainly, the sampling bias is attributed to the limited scale of target data. Noteworthy that the random deviation on unknown categories contributes even more to this bias, as the distribution of unknown categories in an open environment is hard to be retrieved from the sampled unlabeled data at hand.

D. Sampling Bias-Free Upper Bound for SSL Error

Building on Theorem 1, this subsection establishes a sampling bias-free upper bound for SSL error in Theorem 2. Two supporting lemmas used in Theorem 2 are also presented; for detailed proofs, please refer to Appendix IV-D & IV-E.

Lemma 1. *Let $\Phi(\boldsymbol{x}) : \mathcal{X} \to \mathbb{R}^K$ denote a labeling function that is λ^μ-Lipschitz continuous over the input space $\mathcal{X}$ and $\eta_k(\boldsymbol{x}) = p(\varphi(\boldsymbol{y}) = k|\boldsymbol{x})$ be a class-specific regression function that is λ^η-Lipschitz continuous for all k where $\varphi(\boldsymbol{y}) : \mathcal{Y} \to \mathbb{R}$ and $k \in \mathbb{R}$. Additionally, given an instance sampled from ρ_X, there exists a labeled one in its proximity such that their Euclidean distance is bounded by D. Assuming for any $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ within $\mathcal{Y}$, their Euclidean distance remains bounded by L, we obtain that:*

$$\mathbb{E}_{\varphi(\boldsymbol{y}_i)\sim\eta(\boldsymbol{x}_i)}\left[\|\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i\|_2^2\right] \le D^2\left[\lambda^\mu + \lambda^\eta L^2 K\right]. \quad (8)$$

Lemma 2. *Let $\tilde{\omega}$ denote a weight distribution with expectation μ and variance σ^2, and M be a constant. For any $w_i \sim \tilde{\omega}$, the following holds:*

$$\mathbb{E}_{w_i\sim\tilde{\omega}}(\|(w_i - \hat{\epsilon})M\|_2) \le M\sqrt{(\mu - \hat{\epsilon})^2 + \sigma^2}, \quad (9)$$

$$\mathbb{E}_{w_i\sim\tilde{\omega}}(\|w_i M\|_2) \le M\mu. \quad (10)$$

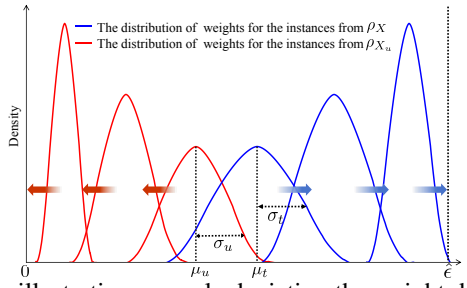


Fig. 6: An illustrative example depicting the weight distribution of instances from ρ_X and ρ_{X_u}.

In the following theorem, we investigate the upper bound of SSL error by leveraging the statistical properties of the true marginal distributions ρ_X and ρ_{X_u}, instead of the empirical marginals $\tilde{\rho}_X$ and $\tilde{\rho}_{X_u}$ used in Theorem 1.

Theorem 2. *(Sampling bias-free upper bound for SSL error). Suppose that $\|w_i - \hat{\epsilon}\|_2$ is bounded by R, where $w_i \sim \omega_t$. Under the same assumptions as presented in Lemma 1, Lemma 2, and Theorem 1, with a probability of at least $1-\delta(\delta>0)$, we can establish the following inequality:*

$$\begin{aligned}&\left|\hat{\mathcal{L}}_P(\hat{h};\boldsymbol{w}) - \hat{\mathcal{L}}_S(h^*;\tilde{\boldsymbol{w}})\right| \\ &\leq \underbrace{\lambda^l \kappa D\sqrt{(\mu_t^2+\sigma_t^2)(\lambda^\mu+\lambda^\eta L^2 K)} + H\sqrt{(\mu_t-\hat{\epsilon})^2+\sigma_t^2} + \mathcal{R}_p}_{\textit{Pseudo-labeling error}} \\ &+ \underbrace{\kappa H \mu_u + \mathcal{R}_i}_{\textit{Invasion error}},\end{aligned} \tag{11}$$

wherein,

$$\mathcal{R}_p = \sqrt{\frac{(\lambda^l \tau_t \kappa L + HR)^2 \log\frac{2}{\delta}}{2s}}, \mathcal{R}_i = \sqrt{\frac{(\tau_u \kappa H)^2 \log\frac{2}{\delta}}{2(n-s)}}.$$

***Remark* 2.** By leveraging the expectations of $\|\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i\|_2^2$ and $\boldsymbol{w}$, as shown in Lemma 1 and Lemma 2, Theorem 2 establishes a sampling bias-free upper bound for SSL error. This bound holds for any set of instances sampled from the true target distribution ρ_X and the true unknown distribution ρ_{X_u}, thereby it is robust for sampling. As the component of SSL error, the bounds of the pseudo-labeling error and the invasion error, are free from the sampling bias. For a detailed proof of Theorem 2, please refer to the Appendix IV-F.

i) Sampling Bias-Free Upper Bound for Pseudo-labeling Error. Theorem 2 states that the sampling bias-free upper bound of the pseudo-labeling error is dominated by the parameters D, μ_t, and σ_t, where D portrays the risk of incorrect annotation, and the others represent the numerical characters of the weight distribution of the instances from ρ_X. Specifically, D denotes the Euclidean distance bound between a labeled instance and an unlabeled one, both obeying the distribution ρ_X. A reduced D implies a closer proximity between labeled and unlabeled instances in Euclidean space. This proximity, as shown in (8), results in a declined expected squared error between the pseudo-label $\hat{\boldsymbol{y}}_i$ and the ground-truth label $\boldsymbol{y}_i$. Consequently, a smaller D, which means less risk for incorrect pseudo-labeling, may mitigate the pseudo-labeling error.

Unlike the parameter D, the paired parameters μ_t and σ_t characterize the weight distribution of the instances following the distribution ρ_X. As the expectation of weights μ_t approaches $\hat{\epsilon}$ and its standard deviation σ_t shrinks to zero in the meantime, the term $H\left(\sqrt{(\mu_t-\hat{\epsilon})^2+\sigma_t^2}\right)$ diminishes. A tighter bound for the pseudo-labeling error is then achieved. Therefore, to mitigate the pseudo-labeling error, it is imperative to align the weights of all instances from the distribution ρ_X, i.e., for any $w_i \sim \omega_t$, with those of the labeled instances, as depicted in Fig. 6.

ii) Sampling Bias-Free Upper bound for Invasion Error. Theorem 2 presents a sampling bias-free upper bound for the invasion error, which manifests the significance of the weights of the instances from the distribution ρ_{X_u}. As demonstrated in (11), the invasion error is controlled by the expectation of the weights, i.e., μ_u. As μ_u tends toward zero, its corresponding variance always converges to zero, owing to $w_i \geq 0$ for any $w_i \sim \omega_u$. In the case that both the expectation and the variance approach zeros, all the weights on instances from ρ_{X_u} approximate zeros, thereby effectively mitigating the invasion of unknown categories.

Summary of Theorem 2. Theorem 2 unveils that the SSL error, including the pseudo-labeling error and the invasion error, is dominant by pseudo-labeling and weight assignment on the basis of the true marginal distributions ρ_X and ρ_{X_u}, instead of the empirical marginal distributions $\tilde{\rho}_X$ and $\tilde{\rho}_{X_u}$. Accordingly, the dual objectives to mitigate the SSL error are pinpointed as improving the pseudo-label quality for the instances from ρ_X and assigning adaptive weights on all unlabeled instances from both ρ_X and ρ_{X_u}. Yet, Theorem 2, together with Theorem 1, is built based on the prior knowledge that the unlabeled instances from ρ_X and the ones from ρ_{X_u} are discernible and divided. However, in the real scenarios, it is impractical to distinguish the unlabeled instances according to their categories, which makes it infeasible to calculate the parameters μ_t, σ_t, and μ_u. Thus, though the bounds in Theorems 1 & 2 are reasonable and explainable, they fail to provide practical guidance for SSL's algorithm design. To address these problems, we present Theorem 3, which considers the unlabeled instances with no prior knowledge.

E. Prior-Free Upper Bound for SSL Error

This subsection establishes an upper bound for the SSL error in Theorem 3, without requiring prior knowledge of the input spaces $\mathcal{X}$ or $\mathcal{X}_u$. To accomplish this, we begin by strengthening the assumptions related to the labeling function, which were previously utilized in Lemma 1, so that they are applicable to the combined input space $\mathcal{X} \cup \mathcal{X}_u$, as depicted in Assumption 1.

Assumption 1. *(Strengthened Assumptions for Lemma 1). Suppose that $\Phi(\boldsymbol{x}) : \mathcal{X} \cup \mathcal{X}_u \to \mathbb{R}^K$ represent a labeling function that exhibits λ^μ-Lipschitz continuity across the input space $\mathcal{X} \cup \mathcal{X}_u$. Additionally, let $\eta_k(\boldsymbol{x}) = p(\varphi(\boldsymbol{y}) = k|\boldsymbol{x})$ denote a class-specific regression function that demonstrates λ^η-Lipschitz continuity for all k, where $\varphi(\boldsymbol{y}) : \mathcal{X} \cup \mathcal{X}_u \to \mathbb{R}$, $k \leq K$, and $k \in \mathbb{R}$. Additionally, for any instance in $\mathcal{X} \cup \mathcal{X}_u$, there exists a labeled one in its proximity such that their Euclidean distance is bounded by D.*

Theorem 3. *(Prior-free upper bound for SSL error). Consider a Lipschitz continuous regression function $\Lambda(\boldsymbol{x}) : \mathcal{X} \cup \mathcal{X}_u \to$*

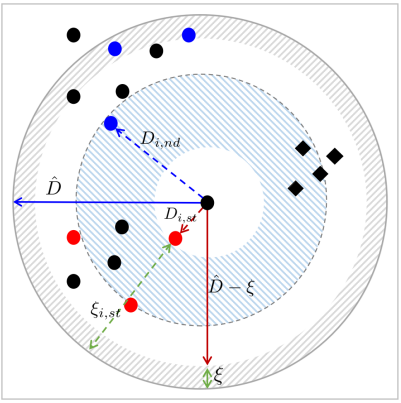


Fig. 7: A toy example for interpreting Theorem 3. Labeled instances with target categories are represented by • and •, while unlabeled instances with target or unknown categories are denoted by • and ⧫, respectively.

$\mathbb{R}$ with constants λ^w on spaces $\mathcal{X}\cup\mathcal{X}_u$. For any unlabeled instance $\boldsymbol{x}_i \in \mathcal{X}\cup\mathcal{X}_u$, there exist labeled instances $\boldsymbol{x}_k$ and $\boldsymbol{x}'_j$ within its proximity, bounded by $\hat{D}$ and $\hat{D}-\xi$ respectively ($\boldsymbol{y}_k \neq \boldsymbol{y}'_j$, $\xi \geq 0$). Under the same assumptions of Assumption 1 and Theorem 2, with a probability of at least $1-\delta(\delta>0)$, the following holds:

$$\left|\hat{\mathcal{L}}_P(\hat{h};\boldsymbol{w}) - \hat{\mathcal{L}}_S(h^*;\tilde{\boldsymbol{w}})\right| \leq \underbrace{(\lambda^l \kappa D\sqrt{\lambda^\mu + \lambda^\eta L^2 K} + 3H)}_{\textit{Pseudo-labeling quality}} \underbrace{(\lambda^w(\hat{D}-\xi)+\hat{\epsilon})}_{\textit{Adaptive weight}} + \mathcal{R}_p + \mathcal{R}_i. \tag{12}$$

***Remark* 3.** Theorems 1 & 2 unveil the dual objective, i.e., improving the pseudo labels and assigning adaptive weights on unlabeled instances, to mitigate the SSL error. However, these results are impractical for SSL algorithm design, as they rely heavily on the prior knowledge of discernible unlabeled instances from unknown categories. To achieve bi-objective optimization, Theorem 3 establishes a prior-free upper bound for the SSL error and outlines several fundamental factors crucial to the quality of pseudo-labeling and adaptive weight assignments. The detailed proof is presented in Appendix IV-G.

As stated in Theorem 3, there are two fundamental factors, D and $\hat{D}-\xi$, decisive to the upper bound of the SSL error. They provide strategies practical to optimize the bi-objective: the quality of pseudo-labeling and adaptive weights.

i) Pseudo-label Quality. Theorem 3 underscores the importance of the parameter D in evaluating the quality of pseudo-labels. Here, D represents the bound for the Euclidean distance between an unlabeled instance $\boldsymbol{x}_i \in \mathcal{X}\cup\mathcal{X}_u$ and a labeled one $\boldsymbol{x}_j$. Among the distances of every unlabeled instance $\boldsymbol{x}_i$ to its nearest labeled neighbor $\boldsymbol{x}_{i,st}$, denoted by $D_{i,st}$, the maximum is assigned to D. Then the presence of a labeled point within the boundary of D for each unlabeled one is ensured, and D reaches its minimum simultaneously. Then the Euclidean distance $D_{i,st}$ may serve as a practical estimator for assessing the pseudo-label quality of each instance. The smaller $D_{i,st}$, the more stringent the penalization on the fundamental factor D, leading to a tighter bound on SSL error.

ii) Adaptive Weights. Theorem 3 unveils that the term $\hat{D}-\xi$ is a fundamental factor on the adaptive weight assignments. A smaller $\hat{D}-\xi$ is conducive to a tighter SSL error bound. To facilitate the following discussion, we visualize the parameters, such as $D_{i,st}$, $D_{i,nd}$, and $\xi_{i,st}$, in Fig. 7.

Obviously, a smaller $\hat{D}$ contributes to a more stringent penalization on the quality $\hat{D}-\xi$, where $\hat{D}$ signifies the Euclidean distance bound between an unlabeled instance $\boldsymbol{x}_i$ and a labeled one $\boldsymbol{x}_k$ with distinct ground-truth labels from $\boldsymbol{x}_j$, just as illustrated in Fig. 7. Considering $\boldsymbol{x}_{i,nd}$ as the closest labeled instance to $\boldsymbol{x}_i$ meeting the condition $\boldsymbol{y}_{i,nd} \neq \boldsymbol{y}_{i,st}$, and $D_{i,nd}$ as the Euclidean distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_{i,nd}$, the inequality $D_{i,nd} \leq \hat{D}$ holds for all $\boldsymbol{x}_i$. Thus, assigning the maximum value among the $D_{i,nd}$ of all instances in $\mathcal{X}\cup\mathcal{X}_u$ to $\hat{D}$ becomes crucial to significantly enhance weight adaptability. This guarantees the presence of two labeled instances with distinct ground-truth labels within the boundary defined by $\hat{D}$ for each unlabeled instance, in the meanwhile $\hat{D}$ reaches its minimum. Upon the above discussion, the Euclidean distance $D_{i,nd}$ is positively correlated with $\hat{D}$, thereby feasible to guide the assignment of weights. Accordingly, $D_{i,nd}$ may serve as a robust estimator for assessing weights in a practical manner.

Contrary to parameter $\hat{D}$, a larger ξ benefits a more rigorous penalty on the quality $\hat{D}-\xi$, where ξ represents the width of the annulus centered at $\boldsymbol{x}_i$ with radii $\hat{D}$ and $\hat{D}-\xi$. For each unlabeled instance $\boldsymbol{x}_i$, we equate $\hat{D}-\xi_{i,st}$ to $D_{i,st}$, where $\xi_{i,st} \geq \xi$ holds for all $\boldsymbol{x}_i$. This guarantees there must be a labeled instance bounded within the $\hat{D}-\xi_{i,st}$ boundary for the corresponding unlabeled instance. To achieve this for the $\hat{D}-\xi$ boundary and maximize ξ, it becomes essential to set ξ as the minimum among $\xi_{i,st}$ for any $\boldsymbol{x}_i$ in $\mathcal{X}\cup\mathcal{X}_u$. Moreover, we conclude that $\xi_{i,st}$ is negatively correlated with $D_{i,st}$ due to $\xi_{i,st} = \hat{D} - D_{i,st}$. To sum up, $\xi_{i,st}$ joint with $D_{i,st}$ may serve as estimators for each instance's weight.

Summary of Theorem 3. Theorem 3 establishes a prior-free upper bound for SSL error and discovers the fundamental factors crucial for improving the quality of pseudo-labels and assigning adaptive weights. Specifically, parameter $D_{i,st}$ may serve as an estimator for pseudo-label quality, while $D_{i,nd}$, $\xi_{i,st}$, and $D_{i,st}$ could serve as estimators for weight assignments. Smaller values of $D_{i,st}$, and larger values of $D_{i,nd}$ and $\xi_{i,st}$ may enhance both pseudo-label quality and weight adaptability. Obviously, these insights benefit specific design strategies on pseudo-labels and adaptive weights.

Summary of BOOM. BOOM comprises three theorems on the upper bound of the SSL error, presented in Subsections III-C, III-D, and III-E. Theorem 1, from an empirical perspective, decouples the SSL error into the pseudo-labeling error and the invasion error and proposes an empirical upper bound. Considering the sampling bias of empirical distribution, Theorem 2 establishes a sampling bias-free upper bound for the SSL error on the true distribution. Both Theorem 1 and Theorem 2 manifest the dual objectives to alleviate the SSL error: high-quality pseudo-labels and adaptive weights. To optimize these bi-objectives, Theorem 3 discovers the fundamental factors that can work as estimators to guide the SSL algorithm design. In a word, by Theorems 1 & 2, BOOM reveals the essence of excess risk of SSL under mismatch, which is explainable and human-understandable. By Theorem 3,

BOOM provides practical guiding principles for algorithmic design in SSL under class distribution mismatch.

IV. Application

By leveraging the estimators unveiled in Theorem 3, we propose the application of BOOM to improve SSL under class distribution mismatch.

Pseudo-label Learning. According to Theorem 3, $D_{i,st}$ may work as an estimator on the quality of the pseudo-labels by measuring the Euclidean distance between an unlabeled instance and its nearest labeled counterpart. To achieve a smaller pseudo-labeling error, a smaller $D_{i,st}$ is preferred. Consequently, we assign the ground-truth label of the nearest labeled instance to the corresponding unlabeled one, as delineated in (13).

$$\hat{\boldsymbol{y}}_{i,u} = \arg\max_{\boldsymbol{k}} f(\boldsymbol{z}_{i,u}, \boldsymbol{z}^{\boldsymbol{k}}_{i,st}), \tag{13}$$

where $f(\boldsymbol{z}_{i,u}, \boldsymbol{z}^{\boldsymbol{k}}_{i,st}) = (2 - D^2_{i,st})/2$ and $D_{i,st} = \|\boldsymbol{z}_{i,u} - \boldsymbol{z}^{\boldsymbol{k}}_{i,st}\|_2$.

Here, $\boldsymbol{z}_{i,u}$ represents the feature learned by contrastive learning for the unlabeled instance, and $\boldsymbol{z}^{\boldsymbol{k}}_{i,st}$ denotes the feature of the nearest labeled one with ground-truth label $\boldsymbol{k}$, where $\|\boldsymbol{z}_{i,u}\|_2 = 1$. (13) quantifies the likeness between the two feature vectors, aiming to identify the ground-truth label $\boldsymbol{k}$ that maximizes this similarity. Note that the unlabeled instances with unknown categories are assigned pseudo-labels by (13) since they cannot be identified with target categories. Additionly, $f(\boldsymbol{z}_{i,u}, \boldsymbol{z}^{\boldsymbol{k}}_{i,st})$ is equivalent to $\cos < \boldsymbol{z}_{i,u}, \boldsymbol{z}^{\boldsymbol{k}}_{i,st} >$ here.

Adaptive Weighting. As suggested by Theorem 3, the estimator on weight adaptability involves $D_{i,nd}$, $\xi_{i,st}$, and $D_{i,st}$. Considering the negative correlation between $\xi_{i,st}$ and the $D_{i,st}$, it is feasible to substitute $\xi_{i,st}$ with $D_{i,st}$. To alleviate the pseudo-labeling error, instances from target categories need larger weights, leading to a preference for a smaller $D_{i,st}$ and a larger $D_{i,nd}$. Conversely, instances from unknown categories are supposed to have smaller weights, resulting in a preference for a larger $D_{i,st}$ and a smaller $D_{i,nd}$, so as to diminish the invasion error. Therefore, the weights are designed as presented in (14).

$$w_{i,u} = g_1\left(f(\boldsymbol{z}_{i,u}, \boldsymbol{z}^{\boldsymbol{k}}_{i,st})\right) \times g_2\left(1 - \frac{f(\boldsymbol{z}_{i,u}, \boldsymbol{z}^{\boldsymbol{v}}_{i,nd})}{f(\boldsymbol{z}_{i,u}, \boldsymbol{z}^{\boldsymbol{k}}_{i,st})}\right), \tag{14}$$

where $g_1(\cdot)$ and $g_2(\cdot)$ represent two monotonically increasing functions, $\boldsymbol{z}^{\boldsymbol{v}}_{i,nd}$ denotes the feature of instance $\boldsymbol{x}_{i,nd}$, $f(\boldsymbol{z}_{i,u}, \boldsymbol{z}^{\boldsymbol{v}}_{i,nd}) = (2 - D^2_{i,nd})/2$, $D_{i,nd} = \|\boldsymbol{z}_{i,u} - \boldsymbol{z}^{\boldsymbol{v}}_{i,nd}\|_2$, and $\boldsymbol{v} \neq \boldsymbol{k}$.

As depicted in (14), $g_1(\cdot)$ exhibits a negative correlation with $D_{i,st}$. A larger weight would be assigned to the instance in case its value of $g_1(\cdot)$ is larger (i.e., $D_{i,st}$ is smaller). This guarantees that the instances with aligned pseudo-labels and ground-truth labels are assigned to larger weights. Moreover, $g_2(\cdot)$ is proportional to $D_{i,nd}$. A small weight is assigned to the instance when its value of $g_2(\cdot)$ is smaller (i.e., $D_{i,nd}$ is smaller). This ensures that a smaller weight is assigned to the instance from the unknown marginal distribution ρ_{X_u}. In a word, the weighting technique depicted in (14), works as a filter for the instances with unknown categories and those incorrectly annotated. Meanwhile, it encourages instances with high-quality pseudo-labels to sufficiently participate in training.

Updating Pseudo-Labels and Weights. To progressively update the pseudo-labels and weights of unlabeled instances, we select some reliable instances to add to the labeled data. By leveraging the dissimilarity between the output of the target classifier and the pseudo-label, we evaluate the reliability of the instance, as formulated in (15).

$$\pi_{i,u}(\hat{h}_{\boldsymbol{\theta}_t}) = \ell(\hat{h}_{\boldsymbol{\theta}_t}(\boldsymbol{x}_{i,u}), \hat{\boldsymbol{y}}_{i,u}) \tag{15}$$

where $\ell(\cdot,\cdot)$ represents the cross-entropy function, and $\boldsymbol{\theta}_t$ denotes the parameters of the target classifier at the t-th iteration.

If $\pi_{i,u}$, as depicted in (15), takes a lower value, then $\boldsymbol{x}_{i,u}$ is considered more reliable. According to this finding, we choose the top $\alpha\%$ reliable instances from the unlabeled data and incorporate them into the labeled data. In the meantime, these selected instances are removed from the unlabeled pool. Additionally, we adopt the polynomial decay [39] to dynamically adjust α to prevent the gradually increasing negative influence from unknown categories with the iteration. The details are shown in Appendix II-A. As a consequence, the updated pseudo-labels and weights progressively improve the target classifier by (3).

Concisely, guided by BOOM, we exploit pseudo-labeling and adaptive weight to mitigate the SSL error. A detailed depiction of this process is outlined in Appendix II-B.

V. Experiments

Subsection V-B presents the comparison results between BOOM and eight state-of-the-art SSL approaches, as well as one standard baseline. Furthermore, an ablation study is conducted in Subsection V-C, while sensitivity analyses and visualization are carried out in Appendix III-A and Appendix III-B, respectively. Additionally, the experimental results on a realistic dataset are shown in Subsection V-D.

A. Experimental Setups

Datasets. Our experiments are conducted on four datasets, including three benchmark datasets, CIFAR10 [40], CIFAR100 [40], and Tiny-Imagenet [41], along with an artificial cross-dataset. The CIFAR10 and CIFAR100 dataset comprises 50,000 training and 10,000 testing images of 10 and 100 categories, respectively. Tiny-Imagenet contains 100,000 training and 10,000 testing images across 200 categories. Moreover, the cross-dataset comprises subsamples from CIFAR10, CIFAR100, Flowers [42], Food-101 [43], and Places-365 [44]. It consists of 138,000 unlabeled instances from 674 categories. All images within these datasets are uniformly resized to 32×32 pixels. For further details, please refer to Appendix III-D.

Settings. i) The proportion of the instances with unknown categories in unlabeled data, named as mismatch proportion, are set as 0%, 20%, 40%, 60%, 80%, and 100% in this work. For example, at a 60% mismatch proportion, the unlabeled data comprises 4,000 instances with target categories and 6,000 instances with unknown categories, while at 0% mismatch proportion, the unlabeled data exclusively contains 4,000 instances with target categories. **ii)** Labeled data is constructed by randomly sampling 8% instances from the training dataset that belong to target categories. The remaining 92% of instances

TABLE II: Training details of BOOM and compared methods.

Configuration	Symbol	Value
Feature encoder	$\phi(\cdot)$	Resnet-18 [46]
Target classifier	$\hat{h}(\cdot)$	WideResnet-28-2 [47]
Optimizer	-	Adam [48]
Learning rate	-	5×10^{-4}
Epochs	-	100
Batch size	-	32
Augmentation techniques	-	Random horizontal flipping, Random translation (up to 2 pixels), Gaussian noise with a standard deviation of 0.15, Global contrast normalization [12]
Function	$g_1(\cdot)$ $g_2(\cdot)$	Identical mappings
Ratio of reliable instances	α	Initial $\alpha = 0.1$, decaying five times until reaching 0.

with target categories and some instances with unknown categories are composed of unlabeled data according to the mismatch proportion. Notably, for a 100% mismatch proportion, all unlabeled instances are from unknown categories. For detailed instance counts in labeled or unlabeled data, please refer to the Appendix III-D.

Training Details. We summarize the training details of BOOM and the compared methods, including the backbone and parameters, in Table II. Both the feature encoder and the target classifier are trained from scratch. The feature encoder is trained using SimCLR [45], adhering to all implementation details of SimCLR. Additionally, global contrast normalization and ZCA normalization—techniques commonly used in pre-treatment [12], [13]—are applied to the CIFAR10 dataset. These configurations are consistently applied across all methods and datasets, unless otherwise stated, with other parameters in the compared methods remaining as reported in the original studies. Each approach is evaluated on each dataset with three independent runs, and mean accuracy and standard deviation are reported.

Compared Methods. BOOM is compared against eight state-of-the-art approaches, which encompass two traditional SSL methods (Pseudo-Labeling [24] and FixMatch [25]) and six methods specifically designed to address class distribution mismatch. These methods are DS^{3}L [12], T2T [14], CC-SSL [15], UASD [13], ORCA [30], and the latest method from the same period, IOMatch [29]. Although FixMatch [25] is designed for class distribution matching, it can also be applied to mismatched scenarios due to its filtering mechanism, which leverages a threshold to filter out pseudo-labeled samples with low confidence. To better demonstrate the role of filtering in SSL, a variant, "FixMatch $\mathbf{w}\backslash\mathbf{o}$ fil.", which excludes the filtering mechanism from FixMatch, is compared. Additionally, we establish a standard supervised learning baseline, denoted as "Labeled Only", which is trained exclusively on limited labeled data using cross-entropy loss. A comparison with this baseline reveals the performance gains that SSL methods achieve from unlabeled data under mismatched scenarios.

Furthermore, T2T and ORCA are performed without pre-training tasks for fairness, denoted as "T2T $\mathbf{w}\backslash\mathbf{o}$ pre." and "ORCA $\mathbf{w}\backslash\mathbf{o}$ pre.". For detailed information on these compared methods, please refer to Section II. Please note that this work does not aim to propose a state-of-the-art method but to provide insights into the compared methods' outstanding or failures in theoretical aspects, and build a strong baseline guided by BOOM.

B. Experimental Results

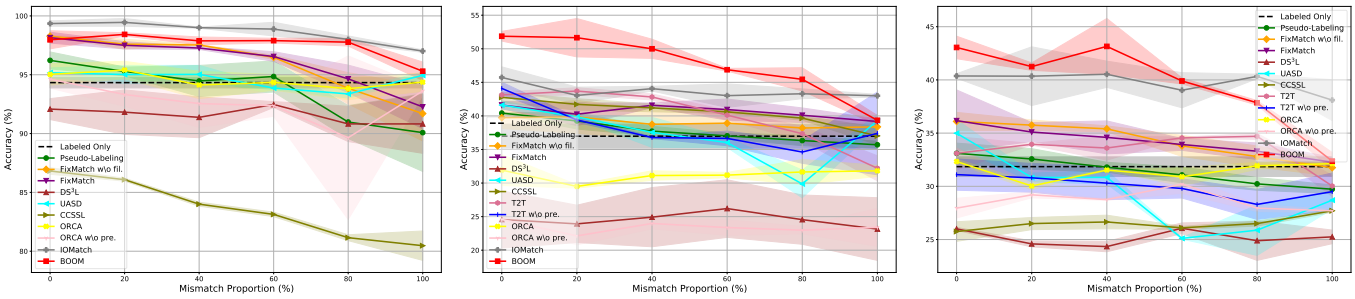
This subsection presents the experimental results from classification tasks conducted across CIFAR10, CIFAR100, Tiny-Imagenet, and a composite cross-dataset, as illustrated in (a), (b), and (c) of Fig. 8 and Fig. 9, respectively. The tables containing specific accuracy values are also reported in Appendix III-C. In CIFAR10, two classes are designated as target categories, while eight remain unknown. The challenge grows significantly in CIFAR100, involving twenty target categories and a substantial eighty unknown categories. Adding to the intricacy, in Tiny-Imagenet, a notable 180 categories fall under the unknown categories, while merely twenty categories are considered targets. Further amplifying the challenge, a cross-dataset is meticulously built, comprising five datasets. In this cross-dataset, six classes from CIFAR10 are specifically chosen as target categories, while an extensive 668 categories from four external datasets are categorically unknown.

Results of mismatched SSL methods. From Fig. 8 and Fig. 9, we have five findings as follows.

i) BOOM demonstrates a significant reduction in both pseudo-labeling error and invasion error. For example, at a 0% mismatch proportion, BOOM (red solid line) shows significant improvements over the "Labeled Only" (black dotted line) across CIFAR10, CIFAR100, Tiny-ImageNet, and the cross-dataset. This highlights the pseudo labeling quality of BOOM. Furthermore, at a 100% mismatch proportion, the accuracy of BOOM surpasses that of "Labeled Only" in most cases.This underscores the effectiveness of the adaptive weight in preventing severe invasion by instances with unknown categories. The moderate increase in performance can be primarily attributed to certain instances with unknown categories yet similarities to the target ones actively contributing to the training process, as indicated in Appendix III-B. Due to the increasing discrepancy between unknown categories and target ones in the cross-dataset, there is a marginal decrease in performance under a 100% mismatch proportion.

ii) BOOM demonstrates its superior performance as a strong baseline, outperforming six compared methods and "Labeled Only" across CIFAR10, CIFAR100, Tiny-Imagenet, and cross-dataset, even when confronted with varying mismatch proportions. In the cross-dataset, although BOOM's accuracy is slightly lower than that of T2T under a 0% mismatch proportion, it outperforms "T2T $\mathbf{w}\backslash\mathbf{o}$ pre.". This suggests that T2T's superior performance is, to some extent, attributed to its pretraining mechanism.

iii) Compared to the robust baseline BOOM, IOMatch demonstrates an unstable ability to mitigate pseudo-labeling error, although it performs better in addressing invasion error. For instance, at 0% mismatch proportions across various datasets, we observe that the accuracy of IOMatch (gray solid line) is lower than that of BOOM on CIFAR100 and Tiny-Imagenet. This instability is attributed to the mechanism for



(a) CIFAR10. (b) CIFAR100. (c) Tiny-Imagenet.

Fig. 8: Experimental results on CIFAR10, CIFAR100, and Tiny-Imagenet under different mismatch proportions, where T2T is not applicable to the binary classification task; thus, accuracy is not reported in (a).

constructing pseudo-labels, i.e., FixMatch, which is sensitive to the count of labeled instances, whereas the mechanism for filtering instances with unknown categories is not. Thus, to enhance SSL performance under class distribution mismatch, it is crucial to explore stable mechanisms to mitigate both pseudo-labeling and invasion errors, as indicated in BOOM.

iv) Some methods show performance decline below the "Labeled Only" under 0% to 100% mismatch proportions. This may be attributed to two factors. First, the invasion error is responsible. It is observed that, in (c) of Fig. 8, UASD (cyan solid line) performs better than the "Labeled Only" at 0% mismatch proportion but declines at non-zero mismatch proportion due to the invasion of instances with unknown categories. Second, the pseudo-labeling error contributes to the decline. We observe that the methods, such as DS^{3}L (brown solid line), CCSSL (olive solid line), and ORCA (yellow solid line), perform lower than the "Labeled Only", even at 0% mismatch proportion. Thus, to enhance the target classifier, it is crucial to address both the pseudo-labeling error and the invasion error simultaneously under class distribution mismatch.

v) Based on the results in Fig. 8 and Fig. 9, it is evident that BOOM exhibits excellent robustness to the scale of unknown categories. For instance, in Tiny-Imagenet and the cross-dataset, BOOM surpasses the "Labeled Only" performance even under an 80% mismatch proportion. It is noteworthy that in the Tiny-Imagenet, the unlabeled data comprises 36,800 instances across 180 unknown categories, while in the cross-dataset, it encompasses a more extensive 110,400 instances spanning 668 unknown categories.

Results of traditional SSL methods under class distribution mismatch. Moreover, we have two observations about traditional SSL methods as follows. **i)** From (a) and (c) of Fig. 8, we observe that the accuracy of traditional SSL methods, such as Pseudo-Labeling (green solid line) and FixMatch (purple solid line), decreases as the mismatch proportion increases, eventually falling below the performance of "Labeled Only" at high mismatch proportions. This decline is attributed to contamination from instances of unknown categories, leading to high invasion error. **ii)** As shown in (a), (b), and (c) of Fig. 8, with mismatch proportions between 20% and 80%, both traditional SSL methods—Pseudo-Labeling and FixMatch—demonstrate lower accuracy than BOOM. This indicates that BOOM effectively mitigates the invasion from unknown categories.

C. Ablation Studies

We conducted ablation studies on the CIFAR10 dataset using various models: "+Pse." (trained with labeled data and unlabeled instances with non-updated pseudo-labels),"+Pse.&W." (trained with non-updated pseudo labels and non-updated weights), and the BOOM model (trained with all components). Additionally, we investigated the individual impact of each component within the weight function and explored various choices for $g(\cdot)$, including identical mapping, $g_i(\cdot)$, and the function $\tilde{g}_i(\cdot) = \exp(\cdot) - 1$.

Effects of pseudo labels. From Fig. 11, we have three findings about the pseudo-labels. **i)** "+Pse." (green solid line) consistently outperforms the "Labeled Only" (black dotted line) across 0% to 80% mismatch proportions. This underscores the efficacy of the pseudo-labeling mechanism within BOOM. **ii)** At a 100% mismatch proportion, "+Pse." falls below the "Labeled Only" as expected. This occurs because all unlabeled instances originate from the unknown marginal distribution ρ_{X_u} and are assigned pseudo-labels from target categories, resulting in an invasion of the target classifier. **iii)** "+Pse." exhibits the comparable performance between 0% and 20% mismatch proportion. The reason is that the unlabeled dataset contain fewer instances with unknown categories under 20% mismatch proportion.

Effects of weights. According to Fig. 11, we observe two findings about weights. **i)** At a 0% mismatch proportion, "+Pse.&W" (orange solid line) achieves comparable accuracy to "+Pse." but demonstrates a reduced standard deviation. This indicates the limited influence of the non-updated weight mechanism in low mismatch proportions. However, its role in improving the target classifier's stability is notable due to it mitigating the impact of potentially incorrect pseudo-labeled instances. **ii)** At a 100% mismatch proportion, "+Pse.&W" remarkably surpasses "+Pse.". This highlights the crucial role of weight in mitigating the negative effects of instances with unknown categories, particularly at a high mismatch proportion. Thus, under class distribution mismatch, target instances with correct pseudo-labels should be assigned larger weights, while instances from unknown categories or those with incorrect pseudo-labels should have weights close to zero.

Effects of updating pseudo-labels and weights. We have two findings about the updating mechanism according to Fig. 11. **i)** The accuracy of BOOM notabely surpasses "+Pse.&W" under various mismatch proportions. This indicates that updating pseudo-labels and weights plays important roles in BOOM. **ii)**

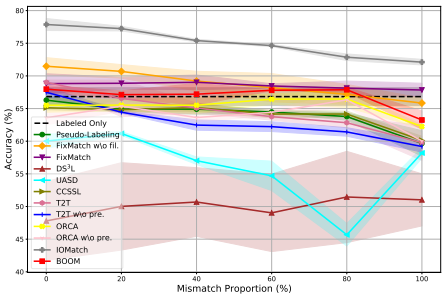


Fig. 9: Results on cross-dataset.

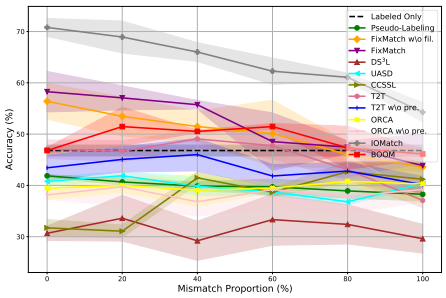


Fig. 10: Results on realistic dataset.

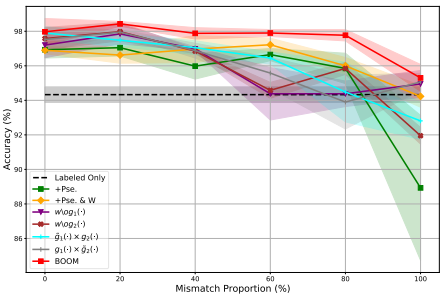


Fig. 11: Ablation studies.

BOOM, training with all components, shows its outstanding performance compared to the ones removing other parts. This demonstrates that the aggregation of all the proposed parts could achieve significant improvement.

Effects of $g_1(\cdot)$ and $g_2(\cdot)$ within weight. From Fig. 11, we have the following four findings. **i)** Both "**w\o** $g_1(\cdot)$" (purple solid line) and "**w\o** $g_2(\cdot)$" (brown solid line) perform worse than BOOM, highlighting their significance for BOOM. **ii)** "**w\o** $g_1(\cdot)$" exhibits a slightly lower performance than "**w\o** $g_2(\cdot)$" at 0% and 20% mismatch proportions. This suggests that when assigned incorrect pseudo-labels to the instances from the marginal distribution ρ_X, $g_1(\cdot)$ demonstrates a greater capacity to alleviate the negative impact than $g_2(\cdot)$. **iii)** However, at 100% mismatch proportion, the performance of "**w\o** $g_1(\cdot)$" notably surpasses that of "**w\o** $g_2(\cdot)$". This emphasizes the stronger ability of $g_2(\cdot)$ to alleviate the adverse effects of instances with unknown categories, illustrating its performance superiority over $g_1(\cdot)$ in such situations. **iv)** The combination $\tilde{g}_1(\cdot)\times g_2(\cdot)$ emphasizes the role of $g_1(\cdot)$, whereas $g_1(\cdot)\times\tilde{g}_2(\cdot)$ highlights the contribution of $g_2(\cdot)$. We observe that $\tilde{g}_1(\cdot)\times g_2(\cdot)$ (cyan solid line) outperforms $g_1(\cdot)\times\tilde{g}_2(\cdot)$ (gray solid line) at 0% mismatch proportion, and conversely at 100% mismatch proportion. This further validates the importance of $g_1(\cdot)$ in mitigating the adverse effects of instances with incorrect pseudo-labels from the marginal distribution ρ_X while demonstrating that $g_2(\cdot)$ excels in filtering instances from unknown categories, as illustrated in iii).

D. Evaluation on Realistic Controlled Noise Dataset

Instead of a synthetic noisy dataset, we evaluate BOOM on a realistic controlled noise dataset, as displayed in Fig. 6 of Appendix III-D. The Red Mini-ImageNet-V2 [1] subset comprises images retrieved from Google Images by Jiang et al. [49], totaling 13,801 instances across 100 categories from Mini-ImageNet. Specifically, five categories ("Triceratops", "Gordon setter", "Alaskan malamute", "Newfoundland dog", "American robin"), each consisting of 600 images, are considered as the targets. For 20% class mismatch proportion, the invaded dataset consists of noise data from these target categories. However, for 40% to 100% class mismatch proportions, the invaded data comprises 100 categories due to insufficient data availability. Moreover, 8% of instances are randomly sampled from the pure instances within target categories to establish the labeled dataset. More details are presented in Appendix III-D.

From Fig. 10, we have the following two findings. **i)** Even on realistic datasets, BOOM significantly outperforms the supervised baseline ("Labeled Only") across mismatch proportions ranging from 0% to 80%, and achieves comparable performance for a 100% mismatch proportion. This underscores its effectiveness as a strong baseline. **ii)** BOOM outperforms the six compared methods due to the slight improvement of T2T over BOOM under a 0% mismatch proportion attributed to the pretraining mechanism. This is evident that BOOM surpasses T2T "**w\o** pre.". Consequently, BOOM demonstrates success even on a realistic dataset, and the theoretical analysis proves effective in the real world.

VI. Conclusions

SSL, as a powerful tool to address the problem of label scarcity, has achieved tremendous progress in applications. However, the conventional SSL methods, assumed on the close-set, typically fall short in overcoming the challenges posed by real-world scenarios, such as class distribution mismatch. Although some SSL approaches have been improved to mismatch settings, the theoretical investigation on generalization analysis of SSL with mismatched class distributions is still under study.

This paper proposes a pioneering theoretical framework for SSL under class distribution mismatch, named BOOM, to evaluate its excess risk. BOOM, composed of three main theorems, reveals that under class distribution mismatch the SSL error predominantly contributes to the excess risk. The first two theorems of BOOM decouple the SSL error into the invasion error and the pseudo-labeling error, and they pinpoint the dual objective (i.e., improving the pseudo-labeling and assigning adaptive weights to unlabeled instances) to mitigate the SSL error. Evidently, these two theorems reveal the essence of the excess risk of SSL under mismatch and they are explainable. To optimize the bi-objectives, Theorem 3 in BOOM profoundly explores the SSL error without any prior knowledge and then discovers several fundamental factors beneficial for SSL algorithm design under mismatch. Guided by the results of BOOM, we propose an SSL method under mismatch which might serve as a strong baseline. Extensive experiments in this paper demonstrate its superior performance. In summary, BOOM bridges the theoretical gap by theoretically interpreting the rationale of SSL under mismatch and provides practical guidelines, such as fundamental factors for bi-objective optimization, for algorithm design to effectively mitigate the SSL error in mismatch settings.

Limitations. In cases that instances with unknown categories share some common features with those target ones, BOOM may struggle to differentiate them within Euclidean space. Indeed, this is a problem faced by all SSL methods under mismatch.

Broader Impact. BOOM provides theoretical guidance for SSL algorithms in mismatched scenarios. More importantly,

it constructs a theoretical framework for excess risk in semi-supervised settings, offering a reference for generalization analysis in weakly supervised learning manners.

VII. Acknowledgement

The authors are grateful to Professor Xizhao Wang, Big Data Institute, Shenzhen University, for his guidance on the decomposition of excess risk in this paper, and to Associate Professor Ran Wang, School of Mathematical Sciences, Shenzhen University, for her valuable assistance with the analysis of version spaces in supervised learning, SSL, and mismatched SSL. This work is supported by the National Key Research & Develop Plan(2023YFB4503603), National Natural Science Foundation of China(U23A20299, 62276270, 62076245, 62072460, 62172424, 62322214), Beijing Natural Science Foundation(4212022), and the Outstanding Innovative Talents Cultivation Funded Programs 2024 of Renmin Univertity of China. It is also partially supported by the Opening Fund of Hebei Key Laboratory of Machine Learning and Computational Intelligence.

References

[1] P. Du, H. Chen, S. Zhao, S. Chai, H. Chen, and C. Li, "Contrastive active learning under class distribution mismatch," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–13, 2022.

[2] C. Wei, K. Shen, Y. Chen, and T. Ma, "Theoretical analysis of self-training with deep networks on unlabeled data," *arXiv preprint arXiv:2010.03622*, 2020.

[3] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009.

[4] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," *Advances in neural information processing systems*, vol. 17, 2004.

[5] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

[6] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[7] P. Rigollet, "Generalization error bounds in semi-supervised classification under the cluster assumption." *Journal of Machine Learning Research*, vol. 8, no. 7, 2007.

[8] A. Singh, R. Nowak, and J. Zhu, "Unlabeled data: Now it helps, now it doesn't," *Advances in neural information processing systems*, vol. 21, 2008.

[9] D. Sanz-Alonso and R. Yang, "Unlabeled data help in graph-based semi-supervised learning: a bayesian nonparametrics perspective," *The Journal of Machine Learning Research*, vol. 23, no. 1, pp. 4205–4232, 2022.

[10] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[11] P. Du, S. Zhao, H. Chen, S. Chai, H. Chen, and C. Li, "Contrastive coding for active learning under class distribution mismatch," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 8927–8936.

[12] L.-Z. Guo, Z.-Y. Zhang, Y. Jiang, Y.-F. Li, and Z.-H. Zhou, "Safe deep semi-supervised learning for unseen-class unlabeled data," in *International Conference on Machine Learning*. PMLR, 2020, pp. 3897–3906.

[13] Y. Chen, X. Zhu, W. Li, and S. Gong, "Semi-supervised learning under class distribution mismatch," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 3569–3576.

[14] J. Huang, C. Fang, W. Chen, Z. Chai, X. Wei, P. Wei, L. Lin, and G. Li, "Trash to treasure: Harvesting ood data with cross-modal matching for open-set semi-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8310–8319.

[15] F. Yang, K. Wu, S. Zhang, G. Jiang, Y. Liu, F. Zheng, W. Zhang, C. Wang, and L. Zeng, "Class-aware contrastive semi-supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 421–14 430.

[16] Y. Fan, A. Kukleva, D. Dai, and B. Schiele, "Ssb: Simple but strong baseline for boosting performance of open-set semi-supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 068–16 078.

[17] Q. Ma, J. Gao, B. Zhan, Y. Guo, J. Zhou, and Y. Wang, "Rethinking safe semi-supervised learning: Transferring the open-set problem to a close-set one," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 370–16 379.

[18] S. Li, S. Ouyang, and Y. Liu, "Understanding the generalization performance of spectral clustering algorithms," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 7, 2023, pp. 8614–8621.

[19] J. Li, B. Wei, Y. Liu, and W. Wang, "Non-iid distributed learning with optimal mixture weights," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2022, pp. 539–554.

[20] P. Du, S. Zhao, Z. Sheng, C. Li, and H. Chen, "Semi-supervised learning via weight-aware distillation under class distribution mismatch," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 410–16 420.

[21] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.

[22] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring," in *International Conference on Learning Representations*, 2019.

[23] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Advances in neural information processing systems*, vol. 32, 2019.

[24] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2, 2013, p. 896.

[25] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.

[26] X. Yang, Z. Song, I. King, and Z. Xu, "A survey on deep semi-supervised learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 9, pp. 8934–8954, 2022.

[27] Y. Wang, P. Qiao, C. Liu, G. Song, X. Zheng, and J. Chen, "Out-of-distributed semantic pruning for robust semi-supervised learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 849–23 858.

[28] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2, 2013.

[29] Z. Li, L. Qi, Y. Shi, and Y. Gao, "Iomatch: Simplifying open-set semi-supervised learning with joint inliers and outliers utilization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 870–15 879.

[30] K. Cao, M. Brbic, and J. Leskovec, "Open-world semi-supervised learning," in *International Conference on Learning Representations*, 2021.

[31] X. Wen, B. Zhao, and X. Qi, "Parametric classification for generalized category discovery: A baseline study," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 590–16 600.

[32] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Generalized category discovery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7492–7501.

[33] S. Ben-David, T. Lu, and D. Pál, "Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning." in *COLT*, 2008, pp. 33–44.

[34] L.-H. Jia, L.-Z. Guo, Z. Zhou, J.-J. Shao, Y. Xiang, and Y.-F. Li, "Bidirectional adaptation for robust semi-supervised learning with inconsistent data distributions," in *International Conference on Machine Learning*. PMLR, 2023, pp. 14 886–14 901.

[35] T. M. Mitchell, "Generalization as search," *Artificial intelligence*, vol. 18, no. 2, pp. 203–226, 1982.

[36] H. Xu and S. Mannor, "Robustness and generalization," *Machine learning*, vol. 86, no. 3, pp. 391–423, 2012.

[37] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *International Conference on Learning Representations*, 2018.
[38] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. USA: Cambridge University Press, 2014.
[39] A. Borichev and Y. Tomilov, "Optimal polynomial decay of functions and operator semigroups," *Mathematische Annalen*, vol. 347, no. 2, pp. 455–478, 2010.
[40] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.
[41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
[42] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1447–1454.
[43] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *European conference on computer vision*. Springer, 2014, pp. 446–461.
[44] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.
[45] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[47] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *British Machine Vision Conference 2016*. British Machine Vision Association, 2016.
[48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.
[49] L. Jiang, D. Huang, M. Liu, and W. Yang, "Beyond synthetic noise: Deep learning on controlled noisy labels," in *International Conference on Machine Learning*. PMLR, 2020, pp. 4804–4815.

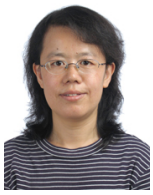
Pan Du is a Ph.D. student at School of Information, Renmin University of China, Beijing, China. She received her BSc degree in Information and Computing Science from Hebei University, Baoding, in 2019, and her MSc degree from Renmin University of China in 2022. Her research interests include machine learning, data mining, and uncertain information processing.

Suyun Zhao received the Bachelor's and Master's degrees from Hebei University, Baoding, China, in 2001 and 2004, respectively, and the Ph.D. degree from The Hong Kong Polytechnic University, Hong Kong. She is currently with Renmin University of China, Beijing, China. Her research interests include machine learning, pattern recognition, and uncertain information processing.

Puhui Tan is pursuing the bachelor degree with the School of Statistics, Renmin University of China, Beijing, China. Her research interests include machine learning, nonparametric statistics, deep learning application.

Zisen Sheng is a three-year MSc student of the College of Information at Renmin University of China, Beijing, China. He received his BSc degree in Data Science and Big Data Technology and Mathematics and Applied Mathematics from Renmin University of China, Beijing, in 2021. His main research interests include machine learning and artificial intelligence.

Zeyu Gan received the bachelor's degree from Renmin University of China, Beijing, China, in 2021. He is currently working toward an MS degree at Renmin University of China. His research interests include weakly-supervised learning, contrastive learning and representation learning.



Hong Chen received the Bachelor and Master degrees from the Renmin University of China, Beijing, China, in 1986 and 1989, respectively, and the Doctor degree from the Chinese Academy of Sciences, Beijing, in 2000. She is currently with Renmin University of China. Her research interests include high-performance database system, data warehouse and data mining, and data management in wireless sensor networks.

Cuiping Li received the B.E. and M.E. degrees from Xi'an Jiao Tong University, Xi'an, China, in 1994 and 1997, respectively, and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2003. She is a Professor with the Renmin University of China, Beijing. Her current research interests include database systems, data warehousing, and data mining.