Collaborative Dual Representations for Semi-Supervised Partial Label Learning

Wei-Xuan Bao[1, 2], **Yong Rui**[3], **Min-Ling Zhang**[1, 2*]

[1]School of Computer Science and Engineering, Southeast University, Nanjing, China
[2]Key Laboratory of Computer Network and Information Integration, Ministry of Education, China
[3]Lenovo Research, Lenovo Group Ltd., Beijing, China
baowx@seu.edu.cn, yongrui@lenovo.com, zhangml@seu.edu.cn

Abstract

Semi-supervised partial label learning (SSPLL) aims to improve the generalization performance of partial label (PL) classifiers by effectively leveraging unlabeled data. Nevertheless, the inherent ambiguity in supervision, where the ground-truth label of a PL example is hidden within a set of candidate labels, poses significant challenges. The presence of false positive labels potentially misleads model's judgment, resulting in pronounced confirmation bias. To address these issues, we propose a novel approach named CoDual, which jointly learns a pair of dual representations for each instance: the predictive class distribution and the low-dimensional embedding. The dual representations interact and progress collaboratively during training. On one hand, in the embedding space the class prototypes are derived via solving a tailored empirical distance minimization problem and employed to smooth the pseudo-targets of unlabeled instances. On the other hand, the refined class distributions regularize the embedding space via encouraging instances with similar pseudo-targets to exhibit similar embeddings. Through an in-depth analysis, we provide-to the best of our knowledge-the first theoretical explanation of how collaborative dual representations facilitate more effective use of unlabeled data for disambiguation. Extensive experiments over benchmark datasets validate the superiority of our proposed approach.

Introduction

Deep learning has made significant strides across various domains. It is widely recognized that data serves as the cornerstone of deep learning methods. Nevertheless, large-scale data annotation in real-world scenarios is often expensive, time-consuming, and labor-intensive. Furthermore, owing to the inherent limitations of annotators' efforts and expertise, the labeling process may inevitably introduce noise, which could have potentially adverse effects on model training. The challenge of ensuring robust generalization performance of learning systems while mitigating the impact of data quantity and quality issues has emerged as a prominent research topic in recent years, commonly referred to as weakly supervised learning (Zhou 2018).

In this study, we focus on a representative weakly supervised learning framework known as partial label learning (PLL) (Cour, Sapp, and Taskar 2011; Bao, Rui, and Zhang 2024). Under this framework, each training instance corresponds to multiple candidate labels, among which only one is valid. The objective of PLL is to learn a multi-class classifier from partial label (PL) training examples to make predictions on unseen instances.

In the context of PLL, it is generally assumed that all training instances are accompanied by a set of candidate labels. Nevertheless, in realistic scenarios, taking medical ultrasound image analysis as an example (Shin et al. 2019), a small amount of experienced experts are only able to annotate a fraction of images, leaving a substantial amount of readily available data unlabeled. Despite their potential to improve the classification performance of the learned model, these unlabeled instances are disregarded in the standard PLL setting. This limitation has led to the emergence of a derived learning framework, referred to as semi-supervised partial label learning (SSPLL) in recent literature (Wang, Li, and Zhou 2019; Wang and Zhang 2020; Song et al. 2022; Wang et al. 2024; Liu et al. 2024; Jiang et al. 2024).

Similar to traditional semi-supervised learning (SSL) (Sun, Shi, and Li 2023), the crucial aspect of addressing the SSPLL problem lies in the effective utilization of unlabeled data. However, it is more challenge for SSPLL since the accessible supervised guidance for unlabeled data is concealed within the candidate label sets of ambiguously labeled examples. The presence of false positive labels could distort model's perception and mislead its judgment, resulting in a pronounced confirmation bias.

In recent research there has been an interesting observation that discriminative models are able to autonomously uncover evident similarities among different categories in the absence of explicit instructions (Wu et al. 2018; He et al. 2020; Huang et al. 2025; Gong and Li 2025). For example, in image classification tasks, the second highest responding class in the model's softmax output is often visually correlated with the given image. In essence, such similarities are not informed by annotations but are intrinsic to the visual data itself. It demonstrates that deep models could actively identify overlapping semantics between categories during the learning process and reflect such similarities in the metric of the output space. This kind of high-order information is valuable and holds great promise as auxiliary supervision to improve model's classification performance.

*Corresponding author

Motivated by the above considerations, in this paper, we propose a novel semi-supervised partial label learning approach named CoDual, i.e. *COllaborative DUAL representations for semi-supervised partial label learning*. In addition to the predictive class probabilities, CoDual also learns an embedding representation for each instance. The two compact representations constitute the dual representations of the same instance, which interact and evolve collaboratively in a holistic learning framework during the training process. Specifically, in the embedding space, CoDual derives the class prototypes via solving a tailored empirical distance minimization problem and utilize them to improve the pseudo-targets of unlabeled instances. Conversely, the refined predictive class probabilities offer more reliable semantic-level correlations between instances, which regularize the structure of embedding space via enforcing examples with similar pseudo-targets to have similar embeddings through a graph reconstruction formulation. We conduct a thorough analysis of our method and provide the first theoretical explanation of how collaborative dual representations enable the model to better leverage unlabeled data for disambiguation. Comprehensive experiments over benchmark datasets validate the superiority of our proposed approach.

The rest of this paper is organized as follows. Section 2 briefly reviews related works on PLL, SSL and SSPLL. Section 3 presents technical details of the proposed approach. Section 4 reports experimental results of extensive comparative studies. Finally, section 5 concludes this paper.

Related Works

Partial Label Learning

As an emerging weakly supervised learning framework, PLL aims to construct a multi-class classifier from training examples with ambiguous labels. In this setting, each instance is associated with a set of candidate labels, while the ground-truth label remains undisclosed and inaccessible throughout the training phase. Intuitively, resolving the ambiguity within candidate labels is the primary approach to addressing PLL challenges. There are two main types of disambiguation strategies, namely identification-based disambiguation strategies and averaging-based disambiguation strategies. For identification-based disambiguation strategies, the unknown ground-truth label is treated as the latent variable and optimized iteratively (Jin and Ghahramani 2002; Liu and Dietterich 2012). For averaging-based disambiguation strategies, candidate labels of PL training examples are treated in the same manner during the training process and the model's outputs are averaged with tailored schemes to yield the final predictions (Cour, Sapp, and Taskar 2011; Tang and Zhang 2017). Several studies also explored operations in the feature space to improve model generalization (Zhang, Wu, and Bao 2022; Bao, Hang, and Zhang 2021, 2022).

In recent years, the deep learning community has increasingly focused on enhancing the model's performance in environments with ambiguous labeling. (Lv et al. 2020) proposes a series of weighted classification losses tailored for PLL and provides theoretical analysis of their consistency and convergency. (Xu et al. 2021) pioneers instance-dependent PLL by employing probabilistic models to iteratively estimate the label distribution for each instance. (Wang et al. 2022) employs iteratively updated class prototypes to construct the contrastive loss, that assists in the discriminative task. (Bao, Rui, and Zhang 2024) introduces disentangled partial label learning, attempting to disentangle the representations of instances and label embeddings to facilitate label disambiguation.

Despite the significant progress made in the field of PLL, existing methods fail to leverage the abundant unlabeled data readily available in real-world scenarios. This represents a major limitation, as these unlabeled data could serve as a new driver for improving the performance of PL learning systems.

Semi-Supervised Learning

Semi-supervised learning represents a well-established field driven by the motivation to leverage the abundance of unlabeled data to improve model's generalization performance (van Engelen and Hoos 2020; Yang et al. 2023). Over decades of research, numerous approaches have been proposed within the SSL framework. Among these, the self-training strategy (McClosky, Charniak, and Johnson 2006; Zou et al. 2018; Xie et al. 2020) constitutes a class of straightforward yet effective approaches. It begins with training a classifier on labeled data, which is subsequently updated by incorporating confident predictions from unlabeled data into the training set. A notable variant of self-training is pseudo-labeling (Lee 2013) where confident model predictions filtered by a threshold are converted into hard labels, implicitly constructing low-entropy pseudo-labels for unlabeled examples. This process has been shown to encourage the classifier's decision boundary to pass through low-density regions of the data distribution (Grandvalet and Bengio 2004). Another foundational technique for SSL is consistency regularization (Bachman, Alsharif, and Precup 2014; Rasmus et al. 2015; Sajjadi, Javanmardi, and Tasdizen 2016), which aims to ensure consistent predictions for different views or perturbations of the same data. Popular approaches to induce the consistency-based loss include data augmentation (Cubuk et al. 2019; Yuan et al. 2021), stochastic regularization (Srivastava et al. 2014; Sajjadi, Javanmardi, and Tasdizen 2016), and adversarial perturbations (Miyato et al. 2019; Jiao et al. 2023).

Nevertheless, conventional SSL methods generally assume that the labeled examples are accurately annotated. When confronted with ambiguously labeled PL examples, the performance of these methods often deteriorates significantly due to the inherent labeling uncertainty.

Semi-Supervised Partial Label Learning

Semi-supervised partial label learning builds on the idea of leveraging unlabeled data to enhance the generalization capability of PL learning systems. This concept was first introduced by (Wang, Li, and Zhou 2019), which employs label propagation to iteratively disambiguate PL examples and assign pseudo-labels to unlabeled instances. Afterwards, (Wang and Zhang 2020) employs the maximum margin formulation to jointly induce the predictive model and estimate labeling confidences over unlabeled training instances. (Song et al. 2022) makes predictions via label set assignment and dependence-maximized dimensionality reduction. (Liu et al. 2024) uniformly addresses both PL examples and unlabeled instances from a mutual information-based perspective. (Jiang et al. 2024) adopts the adaptive threshold to achieve fair selection of confident unlabeled instances and mixes them with PL examples to prevent model overfitting.

In this paper, we propose a novel framework that jointly learns a pair of dual representations for the training data. Notably, we make the first attempt to leverage the spontaneously learned correlations between instances in the output space as explicit guidance for model inference. Comprehensive experiments over benchmark datasets demonstrate that our approach outperforms the state-of-the-art SSPLL algorithms.

The Proposed CoDual Approach

Preliminaries

Notations. Let $\mathcal{X} = \mathbb{R}^d$ represent the d-dimensional input space and $\mathcal{Y} = \{y_1, y_2, ..., y_q\}$ represent the label space with q class labels. Given the PL training set $\mathcal{D}_{\text{PL}} = \{(\mathbf{x}_i, S_i)|1 \leqslant i \leqslant n\}$ and unlabeled training set $\mathcal{D}_{\text{u}} = \{\mathbf{u}_j|1 \leqslant j \leqslant m\}$, where $\mathbf{x}_i = [x_1, ..., x_d]^\top \in \mathcal{X}$, $\mathbf{u}_j = [u_1, ...u_d]^\top \in \mathcal{X}$ are d-dimensional column vectors and $S_i \subseteq \mathcal{Y}$ is the candidate label set associated with the instance $\mathbf{x}_i$, among which only one is the ground-truth label y_i^*, SSPLL aims to derive a multi-class classifier $\hbar : \mathcal{X} \to \mathcal{Y}$ from the training set $\mathcal{D}_{\text{tr}} = \mathcal{D}_{\text{PL}} \cup \mathcal{D}_{\text{u}}$.

Overview. Inspired by recent research observations that discriminative models inherently capture similarities among semantic categories, to address the SSPLL problem, we propose to additionally learn a low-dimensional embedding representation while predicting the class probabilities for each instance. The learned auxiliary embeddings explicitly leverage the instinctively perceived correlations among instances to provide an extra foundation for model inference. This approach establishes two complementary views of the same instance: one from the label space and the other from the feature space, forming a pair of dual representations. The dual representations interact and evolve collaboratively during training, fostering a holistic learning process. Specifically, in the embedding space, the class prototypes are progressively identified and employed to smooth the pseudo-targets for unlabeled instances, leveraging the distance metric based on the Bregman divergence. Conversely, the improved probabilistic predictions offer more accurate guidance for optimizing the embedding space, creating a synergistic feedback loop.

Accordingly, the proposed approach CoDual jointly learns an encoder $f(\cdot)$, a classification head $h(\cdot)$ and a projection head $g(\cdot)$. The encoder $f(\cdot)$ is employed to generate the d^f-dimensional feature vector $f(\mathbf{x}) \in \mathbb{R}^{d^f}$ given the instance $\mathbf{x}$, which serves as the input for the following classification head and the projection head. The classification head $h(\cdot)$ is implemented with a linear model followed by a softmax layer and employed to estimate the distribution over classes $\hat{p}(\mathbf{y}|\mathbf{x})$ for the input $\mathbf{x}$, i.e., the probabilistic vector $\hat{p}(\mathbf{y}|\mathbf{x}) = [\hat{p}(y_1|\mathbf{x}), \hat{p}(y_2|\mathbf{x}), ..., \hat{p}(y_q|\mathbf{x})]^\top = h(f(\mathbf{x}))$. The projection head $g(\cdot)$ is implemented with a non-linear MLP and employed to produce the l_2-normalized embedding vector $z(\mathbf{x}) = \texttt{Normalize}(g(f(\mathbf{x}))) \in \mathbb{R}^{d'}$ for the input $\mathbf{x}$, where d' denotes the dimension of the embedding space. To enhance the exploitation of the ambiguous supervision hidden in PL samples, CoDual introduces the labeling confidence matrix $\mathbf{Y} = [\mathbf{Y}(i,j)]_{n\times q}$, where each element $\mathbf{Y}(i,j)$ denotes the estimated confidence of y_j being the ground-truth label for $\mathbf{x}_i$. This matrix is initialized as Eq.(1):

$$\forall\, 1 \leqslant i \leqslant n,\ 1 \leqslant j \leqslant q:\ \mathbf{Y}(i,j) = \begin{cases} \dfrac{1}{|S_i|}, & \text{if } y_j \in S_i, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

After each epoch, the labeling confidence matrix is re-estimated according to current model's predictive outputs to iteratively disambiguate and progressively reveal the ground-truth labels of PL examples:

$$\mathbf{Y}(i,j) = \begin{cases} \dfrac{\hat{p}(y_j|\texttt{Aug}_{\text{w}}(\mathbf{x}_i))}{\sum_{y_{j'} \in S_i} \hat{p}(y_{j'}|\texttt{Aug}_{\text{w}}(\mathbf{x}_i))}, & \text{if } y_j \in S_i \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where $\texttt{Aug}_{\text{w}}(\cdot)$ denotes a random weak augmentation of the original instance.

Given a batch of PL examples $\mathcal{B}_{\text{PL}} = \{(\mathbf{x}_i, S_i)|1 \leqslant i \leqslant B\} \subset \mathcal{D}_{\text{PL}}$ and a batch of unlabeled instances $\mathcal{B}_{\text{u}} = \{\mathbf{u}_j|1 \leqslant j \leqslant \mu B\} \subset \mathcal{D}_{\text{u}}$, where μ is a hyperparameter which determines the relative sizes of $\mathcal{B}_{\text{PL}}$ and $\mathcal{B}_{\text{u}}$, three types of losses are optimized simultaneously by CoDual, namely the PL classification loss $\mathcal{L}_{\text{PL}}$, the unsupervised classification loss $\mathcal{L}_{\text{u}}$ and the graph reconstruction loss $\mathcal{L}_{\text{re}}$.

For PL examples, the classification loss is defined as the cross-entropy between the disambiguated labeling confidences and the model's predictions:

$$\mathcal{L}_{\text{PL}} = \frac{1}{B}\sum_{i=1}^{B} \text{H}(\mathbf{Y}(i,:), \hat{p}(\mathbf{y}|\texttt{Aug}_{\text{w}}(\mathbf{x}_i))), \tag{3}$$

where $\text{H}(\cdot,\cdot)$ denotes the cross-entropy between two distributions and $\mathbf{Y}(i,:)$ denotes the ith row of the labeling confidence matrix $\mathbf{Y}$.

For unlabeled examples, model's predictions for their weakly-augmented versions $\texttt{Aug}_{\text{w}}(\mathbf{u}_i)(\mathbf{u}_i \in \mathcal{B}_{\text{u}})$ are firstly refined via the class prototype-based pseudo-target smoothing as detailed in section . Consequently, the unsupervised classification loss is defined as the cross-entropy between the smoothed targets $\mathbf{q}_i(1 \leqslant i \leqslant \mu B)$ and the model's outputs for strongly-augmented versions of $\mathbf{u}_i(1 \leqslant i \leqslant \mu B)$:

$$\mathcal{L}_{\text{u}} = \frac{1}{\mu B}\sum_{i=1}^{\mu B} \mathbb{I}(\max(\mathbf{q}_i) \geqslant \tau_q)\text{H}(\mathbf{q}_i, \hat{p}(\mathbf{y}, \texttt{Aug}_{\text{s}}(\mathbf{u}_i))), \tag{4}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function, $\texttt{Aug}_{\text{s}}(\cdot)$ denotes a random strong augmentation of the original instance, and

τ_q is a hyperparameter which helps select trustworthy model predictions as pseudo-targets to further strengthen the training process. [1]

Furthermore, a graph reconstruction-based loss $\mathcal{L}_{\text{re}}$ is designed to leverage the improved pseudo-targets for regularizing the learning of embedding space. The technical details are elaborated in section .

Overall, our training objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{PL}} + \lambda_{\text{u}}\mathcal{L}_{\text{u}} + \lambda_{\text{re}}\mathcal{L}_{\text{re}}, \tag{5}$$

where λ_{u} and λ_{re} are hyperparameters which balance the weights of different loss terms. [2] The pseudo-code of CoDual is provided in Appendix A.

Class Prototype-Based Pseudo-Target Smoothing

Leveraging unlabeled samples to enhance model training is crucial for overcoming the SSPLL problem. Nevertheless, the absence of accurate supervisory information conduces to the unreliability of model judgments. If we adopt the conventional practice in SSL by directly employing the predictive outcomes of weakly augmented unlabeled instances as classification targets for their strongly augmented counterparts, the model is susceptible to suffering from significant confirmation bias, where errors tend to accumulate, thereby adversely impacting the learning process.

Considering the challenges outlined above, CoDual incorporates an additional non-parametric classifier to refine model's original predictive outputs. This classifier is built with a straightforward inductive bias, assuming that instances in the embedding space learned by $g(\cdot)$ naturally cluster around a single prototype representation for each class.

The prototype $\mathbf{s}^*$ could be considered as the representative of a set of N_t data points $\{\mathbf{t}_i\}_{i=1}^{N_t} \subset \mathbb{R}^{d'}$ that belong to the same class, which are encoded by Eq.(6) (Yang et al. 2018):

$$\mathbf{s}^* = \arg\min_{\mathbf{s}} \mathbb{E}_{\mathbf{t}\sim v}[d(\mathbf{t}, \mathbf{s})], \tag{6}$$

where v is a discrete probability distribution over instances $\{\mathbf{t}_i\}_{i=1}^{N_t}$, $d(\cdot,\cdot)$ is the distortion function which measures the divergence between two vectors.

In SSPLL, instances' labeling information remains either concealed or unknown. To address this, CoDual extends Eq.(6) and derives the prototype $\mathbf{s}_j^*$ for each class $y_j \in \mathcal{Y}$ via minimizing the expected distortion between the class prototype and embeddings of the instances associated with that class under the conditional probability $\mathbf{t} \sim p(\mathbf{t}|y_j)$:

$$\mathbf{s}_j^* = \arg\min_{\mathbf{s}} \mathbb{E}_{\mathbf{t}\sim p(\mathbf{t}|y_j)}[d(\mathbf{t}, \mathbf{s})]. \tag{7}$$

Given the excellent properties of the squared Euclidean distance as a regular Bregman divergence (Banerjee et al. 2005), we adopt it as the instantiation of the distortion function $d(\cdot,\cdot)$ in this paper. This choice facilitates the subsequent derivation of class prototypes and the interpretation for the generation of the distributional predictions. Specifically, the regular Bregman divergences denote a particular class of distance functions satisfying the following equation:

$$d_\phi(\mathbf{a}, \mathbf{b}) = \phi(\mathbf{a}) - \phi(\mathbf{b}) - \langle \mathbf{a} - \mathbf{b}, \nabla\phi(\mathbf{b})\rangle, \tag{8}$$

where $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{d'}$, $\phi(\cdot)$ is a differentiable and strictly convex function w.r.t. $d_\phi(\cdot,\cdot)$, $\langle\cdot,\cdot\rangle$ denotes the inner product operation and $\nabla\phi(\cdot)$ denotes the gradient function of $\phi(\cdot)$. Consequently, each class prototype is derived according to Proposition 1, whose proof is provided in Appendix B.

Proposition 1. *Given the bregman divergence $d(\mathbf{a}, \mathbf{b}) = d_\phi(\mathbf{a}, \mathbf{b}) = ||\mathbf{a} - \mathbf{b}||^2$, the prototype of each class $y_j \in \mathcal{Y}$ has a unique formulation to minimize the problem of Eq.(7), which is given by $\mathbf{s}_j^* = \mathbb{E}_{\mathbf{t}\sim p(\mathbf{t}|y_j)}[\mathbf{t}]$.*

Based on the above proposition, the calculation formula for each prototype $\mathbf{s}_j^*(y_j \in \mathcal{Y})$ is derived as follows:

$$\begin{aligned}
\mathbf{s}_j^* &= \mathbb{E}_{p(\mathbf{t}|y_j)}[\mathbf{t}] = \sum_{\mathbf{t}_i\in\mathcal{B}_{\text{PL}}\cup\mathcal{B}_{\text{u}}} p(\mathbf{t}_i|y_j)\mathbf{t}_i \\
&= \sum_{\mathbf{t}_i\in\mathcal{B}_{\text{PL}}\cup\mathcal{B}_{\text{u}}} p(y_j|\mathbf{t}_i)\frac{p(\mathbf{t}_i)}{p(y_j)}\mathbf{t}_i \qquad (9)\\
&= \frac{q}{B}\sum_{\mathbf{x}_i\in\mathcal{B}_{\text{PL}}} \mathbf{Y}(i,j) z(\texttt{Aug}_{\text{w}}(\mathbf{x}_i)) + \\
&\quad \frac{q}{\mu B}\sum_{\mathbf{u}_i\in\mathcal{B}_{\text{u}}} \hat{p}(y_j|\texttt{Aug}_{\text{w}}(\mathbf{u}_i)) z(\texttt{Aug}_{\text{w}}(\mathbf{u}_i)), \qquad (10)
\end{aligned}$$

where Eq.(9) is derived from the Bayes' theorem and Eq.(10) is derived from the assumption that the instances are uniformly distributed and each class has an equal probability of occurrence. Furthermore, the conditional probability $p(y_j|\mathbf{t}_i)$ in Eq.(9) is replaced in Eq.(10) by the estimated labeling confidence for PL examples, and the predictive outputs for unlabeled instances.

After obtaining the prototypes of different classes, CoDual produces a distribution over classes for each unlabeled instance by applying a softmax function over the distances to the prototypes in the embedding space:

$$\bar{p}(y_j|\mathbf{u}_i) = \frac{\exp(-d_\phi(z(\texttt{Aug}_{\text{w}}(\mathbf{u}_i)), \mathbf{s}_j^*))}{\sum_{k=1}^{q}\exp(-d_\phi(z(\texttt{Aug}_{\text{w}}(\mathbf{u}_i)), \mathbf{s}_k^*))}. \tag{11}$$

In this context, we have the following proposition which relates the above straightforward and intuitive probability estimator $\bar{p}(\cdot)$ to learning a mixture of gaussian distributions. The proof is provided in Appendix C.

Proposition 2. *If the distortion function is instantiated with the Bregman divergence $d_\phi(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|^2$, Eq.(11) is equivalent to performing density estimation with the Gaussian Mixture Model (GMM).*

Subsequently, the original predictive outputs of unlabeled instances $\mathbf{u}_i \in \mathcal{B}_{\text{u}}$ are smoothed according to Eq.(12):

$$\mathbf{q}_i = \rho_{\text{s}}\hat{p}(\mathbf{y}|\texttt{Aug}_{\text{w}}(\mathbf{u}_i)) + (1 - \rho_{\text{s}})\bar{p}(\mathbf{y}|\mathbf{u}_i), \tag{12}$$

where $\mathbf{q}_i$ denotes the refined predictive distribution for $\mathbf{u}_i$, ρ_{s} denotes the balancing factor which is set by default as

[1] In this paper, τ_q is set by default to $\tau_q = 0.9$.

[2] In this paper, their default settings are: $\lambda_{\text{u}} = \lambda_{\text{re}} = 1$.

$\rho_s = 0.9$, and $\bar{p}(\mathbf{y}|\mathbf{u}_i) = [\bar{p}(y_1|\mathbf{u}_i), ..., \bar{p}(y_q|\mathbf{u}_i)]^\top$ is the probability vector derived from Eq.(11).

Implementation Details. In this paper, class prototypes are updated progressively after each epoch in an iterative manner. Specifically, the prototype for each category is initially set as a zero vector, i.e., $\mathbf{s}_j^* = \mathbf{0} \in \mathbb{R}^{d'}$ $(y_j \in \mathcal{Y})$. During each epoch, we use the data from the current batch to compute $\hat{\mathbf{s}}_j^*(y_j \in \mathcal{Y})$ according to Eq.(13):

$$\hat{\mathbf{s}}_j^* = \frac{q}{B} \sum_{\mathbf{x}_i \in \mathcal{B}_{\text{PL}}} \mathbf{Y}(i,j) z(\texttt{Aug}_{\text{w}}(\mathbf{x}_i)) + \frac{q}{\mu B} \sum_{\mathbf{u}_i \in \mathcal{B}_{\text{u}}} q_{i,j} \cdot z(\texttt{Aug}_{\text{w}}(\mathbf{u}_i)), \tag{13}$$

which revises Eq.(10) by replacing the original predictive outputs $\hat{p}(y_j|\texttt{Aug}_{\text{w}}(\mathbf{u}_i))$ with the refined probabilities $q_{i,j}$. [3] Then class prototypes are updated with an exponential moving average strategy before being l_2-normalized:

$$\mathbf{s}_j^* \leftarrow \texttt{Normalize}(\rho_{\text{p}} \mathbf{s}_j^* + (1 - \rho_{\text{p}}) \hat{\mathbf{s}}_j^*), \tag{14}$$

where the momentum parameter is set by default to $\rho_{\text{p}} = 0.99$.

Graph Reconstruction-Based Embedding Regularization

In the module of pseudo-target smoothing, CoDual leverages class prototypes derived from the embedding space to refine model's initial predictions. In this module, the smoothed predictive outputs in turn are utilized to regularize the optimization of the embedding space, encouraging the projection head $g(\cdot)$ to produce similar representations for instances with similar pseudo-targets.

Given the smoothed pseudo-targets $\{\mathbf{q}_i\}_{i=1}^{\mu B}$ for the batch of unlabeled instances, CoDual constructs a relationship graph to illustrate the similarities among unlabeled instances from the perspective of labeling semantics. The corresponding affinity matrix $\mathbf{W}^l \in \mathbb{R}^{\mu B \times \mu B}$ is defined as Eq.(15):

$$\mathbf{W}^l(i,j) = \begin{cases} 1 & \text{if } i = j \\ \mathbf{q}_i \cdot \mathbf{q}_j & \text{if } i \neq j \text{ and } \mathbf{q}_i \cdot \mathbf{q}_j \geqslant \tau_g \\ 0 & \text{otherwise,} \end{cases} \tag{15}$$

where only instances with similarity higher than a threshold τ_g are connected in the graph. [4] Furthermore, a self-loop has been incorporated for each node in the graph with the maximum weight $\mathbf{W}^l(i,i) = 1 (1 \leqslant i \leqslant \mu B)$.

The proposed approach CoDual aims to establish consistent relationships of similarities among instances within the embedding space, mirroring those observed in the label space. In order to construct the similarity matrix $\mathbf{W}^f \in \mathbb{R}^{\mu B \times \mu B}$ in the embedding space, two random types of strong augmentations $\texttt{Aug}_{\text{s}}(\cdot)$ and $\texttt{Aug}'_{\text{s}}(\cdot)$ are performed on each unlabeled instances $\mathbf{u}_i \in \mathcal{B}_{\text{u}}$, and the derived embeddings are denoted as $\mathbf{z}_i = z(\texttt{Aug}_{\text{s}}(\mathbf{u}_i))$ and $\mathbf{z}'_i = z(\texttt{Aug}'_{\text{s}}(\mathbf{u}_i))$ for simplicity. Then the embedding graph is defined as:

$$\mathbf{W}^f(i,j) = \begin{cases} \exp(\mathbf{z}_i \cdot \mathbf{z}'_i / \tau_t) & \text{if } i = j \\ \exp(\mathbf{z}_i \cdot \mathbf{z}_j / \tau_t) & \text{if } i \neq j, \end{cases} \tag{16}$$

where the temperature parameter is set by default to $\tau_t = 1$ in this paper.

In this study, two matrices $\mathbf{W}^l$ and $\mathbf{W}^f$ are further normalized so that the sum of elements in each row equals 1:

$$\begin{aligned} \hat{\mathbf{W}}^l(i,j) &= \frac{\mathbf{W}^l(i,j)}{\sum_{k=1}^{\mu B} \mathbf{W}^l(i,k)}, \\ \hat{\mathbf{W}}^f(i,j) &= \frac{\mathbf{W}^f(i,j)}{\sum_{k=1}^{\mu B} \mathbf{W}^f(i,k)}. \end{aligned} \tag{17}$$

Eventually, CoDual minimizes the cross-entropy between the two normalized affinity matrices to recover the graph structure $\hat{\mathbf{W}}^l$ in the embedding space, and the corresponding reconstruction loss is defined as Eq.(18):

$$\mathcal{L}_{\text{re}} = \frac{1}{\mu B} \sum_{i=1}^{\mu B} \text{H}(\hat{\mathbf{W}}^l(i,:), \hat{\mathbf{W}}^f(i,:)). \tag{18}$$

Theoretical Analysis. We undertake an in-depth analysis of the loss function $\mathcal{L}_{\text{re}}$, and identify several advantageous characteristics, which explain why learning dual representations facilitates more effective use of unlabeled data for disambiguation.

Through transforming Eq.(18), we discover that *the minimization of $\mathcal{L}_{re}$ could be viewed as one way of maximizing the mutual information between the embeddings of two strong augmentations $\mathbf{z}_i$, $\mathbf{z}'_i$ of the same instance $\mathbf{u}_i \in \mathcal{B}_u$.* Moreover, if we treat individual instances as distinct classes, as is often the case in unsupervised discriminative tasks, $\mathbf{p}_i = \mathbb{E}_{p(\mathbf{z}|\mathbf{z}_i)}[\mathbf{z}] = \sum_{j=1, j\neq i}^{\mu B} \hat{\mathbf{W}}^l(i,j) \cdot \mathbf{z}_j (1 \leqslant i \leqslant \mu B)$ could be viewed as the prototype for each instance induced by the semantic similarities $\hat{\mathbf{W}}^l$ according to Proposition 1. Then we find that *minimizing $\mathcal{L}_{re}$ could effectively bring each instance $\mathbf{u}_i$ close to its own corresponding prototype $\mathbf{p}_i$ in the embedding space learned by $g(\cdot)$.* For detailed discussions and derivations about the above two properties, please refer to Appendix D.

Furthermore, we derive the gradients of $\mathcal{L}_{\text{re}}$ with respect to the representations $\mathbf{z}_i (1 \leqslant i \leqslant \mu B)$ and discover that *the loss $\mathcal{L}_{re}$ has the intrinsic ability to perform hard negative mining via generating adaptive gradients for optimization.* Detailed derivations are provided in Appendix E.

Experiments

Experimental Setup

Datasets. Four widely used benchmark datasets are employed in our empirical studies to generate the SSPLL datasets, including SVHN (Netzer et al. 2011), CIFAR-10 (Krizhevsky, Hinton et al. 2009), CIFAR-100 (Krizhevsky, Hinton et al. 2009), and STL-10 (Coates, Ng, and Lee 2011). For the SVHN, CIFAR-10 and CIFAR-100 datasets, we first randomly select v samples from each class in the original training set to form a labeled training subset $\mathcal{D}_{\text{l}}$. The remaining training instances, stripped of their labeling information, constitute the unlabeled subset $\mathcal{D}_{\text{u}}$. In our study, the number of labeled instances per class is set to $v \in \{100, 400\}$ for SVHN and CIFAR-10, and $v \in \{50, 100\}$ for CIFAR-100. For STL-10 dataset, since its training set is pre-diveded

[3] The symbol $q_{i,j}$ denotes the jth element of the vector $\mathbf{q}_i$.

[4] In this paper, τ_g is set by default as $\tau_g = 0.8$.

Dataset	Method	$r = 5$	$r = 10$	$r = 15$	$r = 20$
CIFAR-100 ($v = 100$)	CR-DPLL+	62.32 ± 0.67	58.42 ± 0.35	45.32 ± 0.23	43.04 ± 0.54
	TERIAL+	71.08 ± 0.58	64.53 ± 0.46	55.48 ± 0.22	47.18 ± 0.32
	CoMatch+	64.75 ± 0.09	62.12 ± 0.27	57.68 ± 0.31	53.48 ± 0.38
	FreeMatch+	71.81 ± 0.38	69.21 ± 0.34	64.13 ± 0.52	52.08 ± 0.66
	ConCont	71.51 ± 0.84	67.08 ± 0.32	62.95 ± 0.42	53.19 ± 0.48
	SPMI	72.12 ± 0.56	69.71 ± 0.12	64.88 ± 0.37	55.18 ± 0.54
	FairMatch	71.73 ± 0.62	67.96 ± 0.42	64.17 ± 0.82	57.19 ± 0.67
	Ours	$\mathbf{74.67 \pm 0.15}$	$\mathbf{72.85 \pm 0.34}$	$\mathbf{67.96 \pm 0.52}$	$\mathbf{63.84 \pm 0.26}$
CIFAR-100 ($v = 50$)	CR-DPLL+	59.34 ± 0.06	48.26 ± 0.28	44.35 ± 0.65	31.94 ± 0.37
	TERIAL+	57.86 ± 0.23	46.24 ± 0.21	40.38 ± 0.57	29.98 ± 0.49
	CoMatch+	58.59 ± 0.31	51.41 ± 0.43	41.95 ± 0.72	32.88 ± 0.58
	FreeMatch+	65.29 ± 0.51	56.43 ± 0.51	39.84 ± 0.34	27.54 ± 0.31
	ConCont	65.86 ± 0.48	60.12 ± 0.54	48.82 ± 0.43	32.15 ± 0.59
	SPMI	65.98 ± 0.43	58.92 ± 0.46	48.77 ± 0.52	30.13 ± 0.42
	FairMatch	66.28 ± 0.35	59.49 ± 0.93	50.31 ± 0.49	34.66 ± 0.42
	Ours	$\mathbf{69.91 \pm 0.12}$	$\mathbf{64.09 \pm 0.26}$	$\mathbf{54.94 \pm 0.37}$	$\mathbf{43.61 \pm 0.39}$

Table 1: The classification accuracy (mean±std %) of each comparing algorithm on corrupted benchmark dataset of CIFAR-100. The number of labeled instances per class is set to $v \in \{100, 50\}$. The number of false positive labels is set to $r \in \{5, 10, 15, 20\}$. The best results among methods are highlighted in bold.

into labeled and unlabeled subsets, we do not apply further adjustments.

Afterwards, the labeled training subset $\mathcal{D}_{\mathrm{l}}$ is corrupted to form the PL dataset $\mathcal{D}_{\mathrm{PL}}$ following the conventional experimental protocol in PLL (Hüllermeier and Beringer 2006; Cour, Sapp, and Taskar 2011; Gong et al. 2018; Liu and Dietterich 2012). Specifically, r false positive labels are randomly selected for each instance to construct the candidate label set along with the ground-truth label. In our study, the number of false positive labels is set to $r \in \{3, 5, 7\}$ for SVHN, CIFAR-10, STL-10, and $r \in \{5, 10, 15, 20\}$ for CIFAR-100. Accordingly, the comparing models are trained on the combined dataset $\mathcal{D}_{\mathrm{tr}} = \mathcal{D}_{\mathrm{PL}} \cup \mathcal{D}_{\mathrm{u}}$.

Comparing Algorithms. To verify the effectiveness of our proposed approach, CoDual is compared against 3 state-of-the-art SSPLL approaches including ConCont (Wang et al. 2024), SPMI (Liu et al. 2024) and FairMatch (Jiang et al. 2024). Furthermore, following the experimental settings in (Liu et al. 2024), we adapt several advanced PLL and SSL algorithms to the SSPLL problem for comparison. For PLL algorithms, we select CR-DPLL (Wu, Wang, and Zhang 2022) and TERIAL (Bao, Rui, and Zhang 2024) as baselines. Since they could only deal with PL instances, we integrate them with the flexible and effective SSL algorithm FixMatch (Sohn et al. 2020) to enable them to process unlabeled data. Similarly, for SSL algorithms, we select CoMatch (Li, Xiong, and Hoi 2021) and FreeMatch (Wang et al. 2023) as baselines and combine them with the PRODEN loss (Lv et al. 2020) to address the PL data. For clarity, in the following reports, we denote the enhanced version of an algorithm $\mathcal{A}$ as $\mathcal{A}$+.

In this paper, we employ the Wide ResNet-28-2 (Zagoruyko and Komodakis 2016) as the backbone of all deep models. For CoDual, the classification head is a linear model while the projection head is a 2-layer MLP which outputs 64-dimensional embeddings. All deep models are implemented with PyTorch (Paszke et al. 2019) and trained with stochastic gradient descent (SGD) (Robbins and Monro 1951) optimizer with momentum 0.9 on 1 NVIDIA Tesla V100 GPU (32GB). For PLL comparing algorithms, we search the initial learning rate from $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ and the weight decay from $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. For other algorithms, the learning rate is initially set as 0.03 and adjusted with a cosine decay schedule, with the weight decay fixed as 5×10^{-4}. The hyperparameters of all comparing algorithms are specified according to the recommended parameter settings in their respective literature. For CoDual, its hyperparameters are set as previously described. We set the mini-batch size $B = 64$ and the coefficient $\mu = 7$. The training epoch is set as 500. All experiments are conducted five times with different random seeds, and the average accuracy and the standard deviation are reported.

Augmentations. The proposed CoDual involves one weak augmentation $\mathtt{Aug}_{\mathrm{w}}(\cdot)$ and two strong augmentations $\mathtt{Aug}_{\mathrm{s}}(\cdot)$ and $\mathtt{Aug}'_{\mathrm{s}}(\cdot)$. Specifically, the weak augmentation is implemented using the standard crop-and-flip. For strong augmentations, $\mathtt{Aug}_{\mathrm{s}}(\cdot)$ employs RandAugment (Cubuk et al. 2020) following (Sohn et al. 2020) and $\mathtt{Aug}'_{\mathrm{s}}(\cdot)$ applies random color jittering and grayscale conversion as inspired by (Chen et al. 2020).

Method	CIFAR-10		CIFAR-100	
	$v = 400$	$v = 100$	$v = 100$	
	$r = 5$		$r = 10$	$r = 20$
CoDual	94.54 ± 0.27	60.74 ± 0.43	72.85 ± 0.34	63.84 ± 0.26
CoDual w/o PS	93.47 ± 0.43	32.88 ± 0.51	63.38 ± 0.27	54.34 ± 0.19
CoDual w/o GR	92.03 ± 0.35	40.71 ± 0.28	64.88 ± 0.28	49.71 ± 0.41
CoDual w/o PS & GR	91.87 ± 0.22	40.82 ± 0.31	64.72 ± 0.15	49.84 ± 0.43

Table 2: Predictive results (mean±std %) of CoDual and its variants in ablation studies.

Experimental Results

The classification results (mean±std %) of comparing algorithms on SVHN and CIFAR-10 datasets are reported in Table 3 in Appendix F. The experimental results on the CIFAR-100 dataset are reported in Table 1. The experimental results on the STL-10 dataset are reported in Table 4 in Appendix F. In these tables, the best results are highlighted in bold.

In a total of 23 experimental settings, (2 datasets $\times$ 2 settings of $v \times 3$ settings of $r + 1$ dataset $\times$ 2 settings of $v \times$ 4 settings of $r + 1$ dataset $\times$ 1 setting of $v \times 3$ settings of r), CoDual achieves the best performance in 21 cases compared against baseline methods. The only two exceptions occur on datasets of SVHN and STL-10 when the number of false positive labels is set as the minimum value $r = 3$. As the number of available labeled examples decreases and the number of false positive labels increases, the performance advantage of CoDual becomes increasingly pronounced. These impressive results highlight the effectiveness of the reciprocal interaction between dual representations in enabling the proposed algorithm to better disambiguate candidate labels and fully exploit unlabeled data.

Further Analysis

Ablation Studies. To thoroughly evaluate the effectiveness of our proposed SSPLL algorithm, we conduct extensive ablation studies to analyze the contribution of individual components. In this paper, we learn a pair of dual representations within the label space and embedding space for the input data. We facilitate their mutual interaction throughout the learning process, thereby enhancing the quality of both learned low-dimensional representations and the classification results. From the perspective of algorithm implementation, the pseudo-target smoothing module (PS) improves the pseudo-targets of unlabeled instances based on the metric of embedding space, while the graph reconstruction (GR) module employs refined predictive class probabilities to guide the optimization of embedding space. The ablation experiments corresponding to these two fundamental modules are performed on CIFAR-10 ($v \in \{400, 100\}, r = 5$) and CIFAR-100 ($v = 100, r \in \{10, 20\}$) datasets and the results are reported in Table 2. We remove the PS module via setting $\rho_{\text{s}} = 1$ in Eq.(12) and remove the GR module via setting $\lambda_{\text{re}} = 0$ in Eq.(5). It is clear from the table that both of the two fundamental modules significantly enhance the model performance. It is worth noting that the performance improvement is even pronounced under challenging circumstances of insufficient labeled training examples and high rate of false positive labels, which validates the critical importance of our designs in achieving state-of-the-art performance in SSPLL.

Due to page limitations, the analyses of **model parameter sensitivity** and **scalability** are presented in Appendix G and Appendix H, respectively.

Conclusion

In this paper, we propose a novel and theoretically grounded SSPLL algorithm CoDual. It learns a pair of dual representations for instances, explicitly leveraging the spontaneously learned correlations between examples in the output space to guide model inference. Specifically, on one hand, a tailored empirical distance minimization problem is formulated to induce the class prototypes in the embedding space which facilitate the smoothness of the pseudo-targets for unlabeled instances. On the other hand, refined semantic similarities between instances are employed to regularize the optimization of the embedding space. Comprehensive experiments against state-of-the-art SSPLL algorithms show the superiority of our proposed approach.

Acknowledgments

The authors wish to thank the anonymous reviewers for their helpful comments and suggestions. This work was supported by the National Science Foundation of China (62225602), the Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX24_0414) and the Big Data Computing Center of Southeast University.

References

Bachman, P.; Alsharif, O.; and Precup, D. 2014. Learning with pseudo-ensembles. In *Advances in Neural Information Processing Systems 27*, 3365–3373. Montreal, Canada.

Banerjee, A.; Merugu, S.; Dhillon, I. S.; and Ghosh, J. 2005. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6: 1705–1749.

Bao, W.; Rui, Y.; and Zhang, M. 2024. Disentangled partial label learning. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, 11007–11015. Vancouver, Canada.

Bao, W.-X.; Hang, J.-Y.; and Zhang, M.-L. 2021. Partial label dimensionality reduction via confidence-based dependence maximization. In *Proceedings of the 27th ACM*

SIGKDD Conference on Knowledge Discovery and Data Mining, 46–54. Virtual Event.

Bao, W.-X.; Hang, J.-Y.; and Zhang, M.-L. 2022. Submodular feature selection for partial label learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 26–34. Washington D.C.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. E. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, 1597–1607. Virtual Event.

Coates, A.; Ng, A. Y.; and Lee, H. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, volume 15, 215–223. Fort Lauderdale, FL.

Cour, T.; Sapp, B.; and Taskar, B. 2011. Learning from partial labels. *Journal of Machine Learning Research*, 12: 1501–1536.

Cubuk, E. D.; Zoph, B.; Mané, D.; Vasudevan, V.; and Le, Q. V. 2019. AutoAugment: Learning augmentation strategies from data. In *Proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 113–123. Long Beach, CA.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. 2020. RandAugment: Practical automated data augmentation with a reduced search space. In *Advances in Neural Information Processing Systems 33*, 6–12. Virtual Event.

Gong, C.; and Li, X. 2025. Agents with foundation models: Advance and vision. *Frontiers of Computer Science*, 19(4): 194330.

Gong, C.; Liu, T.; Tang, Y.; Yang, J.; Yang, J.; and Tao, D. 2018. A regularization approach for instance-based superset label learning. *IEEE Transactions on Cybernetics*, 48(3): 967–978.

Grandvalet, Y.; and Bengio, Y. 2004. Semi-supervised learning by entropy minimization. In *Advances in Neural Information Processing Systems 17*, 529–536. Vancouver, Canada.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. B. 2020. Momentum contrast for unsupervised visual representation Learning. In *Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9726–9735. Seattle, WA.

Huang, W.; Ye, M.; Shi, Z.; Li, H.; and Du, B. 2025. Self-knowledge distillation with dimensional history knowledge. *Science China Information Sciences*, 68(9).

Hüllermeier, E.; and Beringer, J. 2006. Learning from ambiguously labeled examples. *Intelligent Data Analysis*, 10(5): 419–439.

Jiang, J.; Jia, Y.; Liu, H.; and Hou, J. 2024. FairMatch: Promoting partial label learning by unlabeled samples. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1269–1278. Barcelona, Spain.

Jiao, R.; Liu, X.; Sato, T.; Chen, Q. A.; and Zhu, Q. 2023. Semi-supervised semantics-guided adversarial training for robust trajectory prediction. In *Proceedings of the 19th IEEE/CVF International Conference on Computer Vision*, 8207–8217. Paris, France.

Jin, R.; and Ghahramani, Z. 2002. Learning with multiple labels. In *Advances in neural information processing systems 15*, 897–904. Vancouver, Canada.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. *Technique Report*.

Lee, D.-H. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop of the 30th International Conference on Machine Learning*, volume 3, 896–904. Atlanta, GA.

Li, J.; Xiong, C.; and Hoi, S. C. H. 2021. CoMatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the 18th IEEE/CVF International Conference on Computer Vision*, 9455–9464. Montreal, Canada.

Liu, L.; and Dietterich, T. 2012. A conditional multinomial mixture model for superset label learning. In *Advances in Neural Information Processing Systems 25*, 557–565. Cambridge, MA.

Liu, Y.; Lv, J.; Geng, X.; and Xu, N. 2024. Learning with partial-label and unlabeled Data: A uniform treatment for supervision redundathncy and insufficiency. In *Proceedings of the 41st International Conference on Machine Learning*, 31614–31628. Vienna, Austria.

Lv, J.; Xu, M.; Feng, L.; Niu, G.; Geng, X.; and Sugiyama, M. 2020. Progressive identification of true labels for partial-label learning. In *Proceedings of the 37th International Conference on Machine Learning*, 6500–6510. Virtual Event.

McClosky, D.; Charniak, E.; and Johnson, M. 2006. Effective self-training for parsing. In *Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. New York, NY.

Miyato, T.; Maeda, S.; Koyama, M.; and Ishii, S. 2019. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8): 1979–1993.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning. In *Proceedings of the NIPS Workshop on Deep Learning and Unsupervised Feature Learning*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E. Z.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, 8024–8035. Vancouver, Canada.

Rasmus, A.; Berglund, M.; Honkala, M.; Valpola, H.; and Raiko, T. 2015. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems 28*, 3546–3554. Montreal, Canada.

Robbins, H.; and Monro, S. 1951. A stochastic approximation method. *The Annals of Mathematical Statistics*, 400–407.

Sajjadi, M.; Javanmardi, M.; and Tasdizen, T. 2016. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in Neural Information Processing Systems 29*, 1163–1171. Barcelona, Spain.

Shin, S. Y.; Lee, S.; Yun, I. D.; Kim, S. M.; and Lee, K. M. 2019. Joint weakly and semi-supervised deep learning for localization and classification of masses in breast ultrasound images. *IEEE Transactions on Medical Imaging*, 38(3): 762–774.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C.; Cubuk, E. D.; Kurakin, A.; and Li, C. 2020. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems 33*, 596–608. Virtual Event.

Song, M.; Li, H.; Sun, C.; Cai, D.; and Hong, S. 2022. Dlsa: Semi-supervised partial label learning via dependence-maximized label set assignment. *Information Sciences*, 609: 1169–1180.

Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1): 1929–1958.

Sun, Y.; Shi, Z.; and Li, Y. 2023. A graph-theoretic framework for understanding open-world semi-supervised learning. In *Advances in Neural Information Processing Systems 36*, 23934–23967. New Orleans, LA.

Tang, C.-Z.; and Zhang, M.-L. 2017. Confidence-rated discriminative partial label learning. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, 2611–2617. San Francisco, CA.

van Engelen, J. E.; and Hoos, H. H. 2020. A survey on semi-supervised learning. *Machine Learning*, 109(2): 373–440.

Wang, H.; Xiao, R.; Li, Y.; Feng, L.; Niu, G.; Chen, G.; and Zhao, J. 2022. PiCO: Contrastive label disambiguation for partial label learning. In *Proceedings of the 10th International Conference on Learning Representations*. Virtual Event.

Wang, Q.; Li, Y.; and Zhou, Z. 2019. Partial label learning with unlabeled data. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 3755–3761. Macao, China.

Wang, Q.; Zhao, B.; Zhu, M.; Li, T.; Liu, Z.; and Xia, S. 2024. Controller-guided partial label consistency regularization with unlabeled data. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, 15571–15579. Vancouver, Canada.

Wang, W.; and Zhang, M. 2020. Semi-supervised partial label learning via confidence-rated margin maximization. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 6982–6993. Virtual Event.

Wang, Y.; Chen, H.; Heng, Q.; Hou, W.; Fan, Y.; Wu, Z.; Wang, J.; Savvides, M.; Shinozaki, T.; Raj, B.; Schiele, B.; and Xie, X. 2023. FreeMatch: Self-adaptive thresholding for semi-supervised learning. In *Proceedings of the 11th International Conference on Learning Representations*. Kigali, Rwanda.

Wu, D.-D.; Wang, D.-B.; and Zhang, M.-L. 2022. Revisiting consistency regularization for deep partial label learning. In *Proceedings of the 39th International Conference on Machine Learning*, 24212–24225. Baltimore, MD.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the 29th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3733–3742. Salt Lake City, UT.

Xie, Q.; Luong, M.; Hovy, E. H.; and Le, Q. V. 2020. Self-training with noisy student improves imageNet classification. In *Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695. Seattle, WA.

Xu, N.; Qiao, C.; Geng, X.; and Zhang, M.-L. 2021. Instance-dependent partial label learning. In *Advances in Neural Information Processing Systems 34*, 27119–27130. Virtual Event.

Yang, H.; Zhang, X.; Yin, F.; and Liu, C. 2018. Robust classification with convolutional prototype learning. In *Proceedings of the 29th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3474–3482. Salt Lake City, UT.

Yang, X.; Song, Z.; King, I.; and Xu, Z. 2023. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9): 8934–8954.

Yuan, J.; Liu, Y.; Shen, C.; Wang, Z.; and Li, H. 2021. A simple baseline for semi-supervised semantic segmentation with strong data augmentation. In *Proceedings of the 18th IEEE/CVF International Conference on Computer Vision*, 8229–8238. Montreal, Canada.

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. In *Proceedings of the 27th British Machine Vision Conference*, 19–22. York, UK.

Zhang, M.-L.; Wu, J.-H.; and Bao, W.-X. 2022. Disambiguation enabled linear discriminant analysis for partial label dimensionality reduction. *ACM Transactions on Knowledge Discovery from Data*, 16(4): 72:1–72:18.

Zhou, Z.-H. 2018. A brief introduction to weakly supervised learning. *National Science Review*, 5(1): 44–53.

Zou, Y.; Yu, Z.; Kumar, B. V. K. V.; and Wang, J. 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the 15th European Conference on Computer Vision*, volume 11207, 297–313. Munich, Germany.