Addressing Multi-Label Learning with Partial Labels: From Sample Selection to Label Selection

Gengyu Lyu[1,2,4], **Bohang Sun**[1], **Xiang Deng**[2,3], **Songhe Feng**[*2,5]
[1] College of Computer Science, Beijing University of Technology
[2] School of Computer Science and Technology, Beijing Jiaotong University
[3] Department of Automation, Tsinghua University
[4] Idealism Beijing Technology Co., Ltd.
[5] Key Laboratory of Big Data & Artificial Intelligence in Transportation (Beijing Jiaotong University), Ministry of Education
lyugengyu@gmail.com, sunbohang@emails.bjut.edu.cn, 20120346@bjtu.edu.cn, shfeng@bjtu.edu.cn

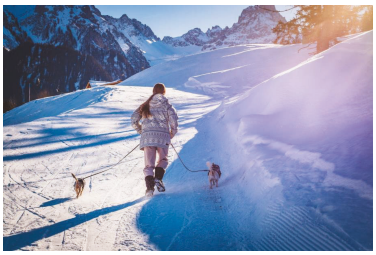
Abstract

Multi-label Learning with Partial Labels (ML-PL) learns from training data, where each sample is annotated with part of positive labels while leaving the rest of positive labels unannotated. Existing methods mainly focus on extending multi-label losses to estimate unannotated labels, further inducing a missing-robust network. However, training with single network could lead to confirmation bias (i.e., the model tends to confirm its mistakes). To tackle this issue, we propose a novel learning paradigm termed Co-Label Selection (CLS), where two networks feed forward all data and cooperate in a co-training manner for critical label selection. Different from traditional co-training based methods that networks select confident samples for each other, we start from a new perspective that two networks are encouraged to remove false-negative labels while keep training samples reserved. Meanwhile, considering the extreme positive-negative label imbalance in ML-PL that leads the model to focus on negative labels, we enforce the model to concentrate on positive labels by abandoning non-informative negative labels to alleviate such issue. By shifting the cooperation strategy from "Sample Selection" to "Label Selection", CLS avoids directly dropping samples and reserves training data in most extent, thus enhancing the utilization of supervised signals and the generalization of the learning model. Empirical results on various multi-label datasets demonstrate that our CLS is significantly superior to other state-of-the-art methods.

Introduction

Multi-label learning aims to predict a set of labels related to a sample, which has been widely applied in a variety of fields ranging from document classification (Liu et al. 2021), image annotation (Wang et al. 2022), social tag recommendation (Vu et al. 2020) and emotion recognition of internet-based multimedia data (Wang et al. 2015). Current studies strongly rely on high-quality fully labelled multi-label datasets with complete and accurate labels, while it is almost impossible to annotate large-scale benchmarks, especially when the number of categories is large. Considering the significant effort required to collect an adequate and exhaustive list of labels for each instance, it is appealing if model can learn from a few available labels and automatically distinguish false-negative and true-negative labels. Following this, (Durand, Mehrasa, and Mori 2019) conduct the first attempt to empirically compare various labelling methodologies to demonstrate the possibility for employing partial labels on multi-label datasets. (Huynh and Elhamifar 2020) regularise the cross-entropy loss with a cost function that gauges the smoothness of labels and sample features characteristics to avoid the overfitting problem in ML-PL. (Cole et al. 2021) extend multi-label losses to handle the extreme case of ML-PL, where annotators only provide one relevant label for each instance. Although these techniques greatly advance the learning of missing-robust network by extending various loss function variations, all of them are formulated with a single network, which inevitably results in confirmation bias since the model tends to corroborate the erroneous data (Tarvainen and Valpola 2017).

In order to combat such confirmation bias problem, co-training based approaches have recently received a lot of attention. (Han et al. 2018) suggest a co-teaching technique that trains two models concurrently, which allows them to choose small loss instances for one another. (Yu et al. 2019) adopt the "Update by Disagreement" strategy to ensure the divergence between two models, which only selects prediction disagreement samples for further co-teaching. (Wei et al. 2020) propose a joint training with co-regularization method named JoCoR, which increases the effective num-



	[a]	[b]	[c]
dog	✓	✓	✓
person	✓	?	×
tree	✓	?	×
mountain	✓	✓	✓
river	×	?	✓
car	×	?	×
horse	×	?	×

Figure 1: Example of sample with ground-truth labels [a], partial label [b], and noisy labels [c]. Samples in ML-PL are annotated with a subset of positive labels and needs less annotation cost than multi-label learning, and can be viewed as a special case of noisy labels.

*Corresponding Author.

ber of co-training samples by formulating a joint loss with an explicit regularization constraint. However, faced with ML-PL problem, the above methods still face two major challenges: 1) The effective number of co-training samples is still limited, especially whether these selected samples have completely ground-truth labels also cannot be guaranteed in ML-PL. 2) During the co-training process, only high confident (low loss) samples are considered while low confident (high loss) samples are directly dropped, which leaves the latent valuable information unused since samples with partial labels tend to result in high loss in ML-PL.

To tackle the above issues, we propose a new perspective of co-training based method termed Co-Label Selection (CLS), which converts the cooperation strategy between two networks from "Sample Selection" to "Label Selection", thereby addressing the confirmation bias in self-training while keeping training samples reserved in great extent. Specifically, we train two networks simultaneously to feed forward and generate the label confidence value for each training sample. Then, each network identifies false-negative labels for its peer network according to the rankings of label confidence. Meanwhile, the non-informative negative labels are also dropped by network itself to avoid the learning model excessively focusing on negative labels. Such operation can help to alleviate the well-known positive-negative label imbalance issue in multi-label learning. According to the above operations, we maintain two networks cooperate at label-level within a mini-batch to generate proper supervision signals for each other, which enhances the utilization of training data by label selection instead of sample dropping operation. Notably, we operate label selection according to the rankings of label confidence within all labels contained in a mini-batch, thus implicitly modeling the label dependency and further improving the prediction accuracy of our proposed method. The main contributions of this paper are summarized as follows:

- We propose a novel co-training based ML-PL method termed CLS, which for the first time converts the cooperation strategy between two networks from "Sample Selection" to "Label Selection" to avoid directly dropping samples and reserve training data to the utmost degree.
- We denote that single "co-training" strategy (e.g., the co-teaching algorithm or JoCoR algorithm) *cannot* handle ML-PL data well, which suffers from the problem of limited effective training samples. Such argument has been empirically justified in Section 4.
- We thoroughly evaluate our proposed method through experiments on various datasets. And extensive results show that CLS achieves promising performances in comparison with other state-of-the-art methods.

Related Work

Multi-label Learning with Partial Labels

Multi-label learning methods (Lyu et al. 2024a,b; Zhong, Lyu, and Yang 2024; Wang et al. 2023, 2024) usually rely heavily on high-quality labelled large-scale training data, while in practice it is expensive and infeasible to obtain such ideal data. On the contrary, it is easy to obtain a subset of labels for training, which is called as Multi-label Learning with Partial Labels (Liu et al. 2023; Gu et al. 2023; Wu et al. 2022; Jia and Zhang 2023; Liu, Jia, and Zhang 2023). To explore how to learn with incomplete annotation, (Sun, Zhang, and Zhou 2010) construct a similarity graph for each label and the manifold regularization term is added to recover the missing labels. (Durand, Mehrasa, and Mori 2019) introduce a normalized binary cross-entropy (BCE) loss that exploits the proportion of known labels and use it to train the model with partial label. (Huynh and Elhamifar 2020) introduce a new loss function that regularizes the smoothness of labels and image features to avoid overfitting to partial label. (Chen et al. 2021a) explore semantic correlations to transfer knowledge of known labels to generate pseudo labels for unknown labels and use both known and generated labels for model training. Although these methods have achieved competitive performance, they learn from partial labeled data by training with a single model, which could fall into confirmation bias (i.e. the model is prone to confirm its mistakes) (Tarvainen and Valpola 2017). Differently, our CLS learns two model simultaneously and let them select proper annotations for each other, which eliminates the negative influence of confirmation bias and improves model performance.

Co-training with Noisy Labels

Single-network based noisy label learning methods usually suffer from the confirmation bias problem (Tarvainen and Valpola 2017) and they tend to accumulate the correction error during training, resulting in low model performance (Feng, An, and He 2019; Min et al. 2023; Ge et al. 2022). To avoid accumulating confirmation bias, researchers propose to train two networks simultaneously to filter errors in a co-training manner. For example, (Han et al. 2018) let two models back propagate data with clean labels to its peer model. (Yu et al. 2019) select small-loss instances with different predictions from two networks to ensure model diversity. (Li, Socher, and Hoi 2020) leverage noisy samples by modeling the per-sample loss distribution to divide the training data to labeled and unlabeled data, and exploit all samples in a semi-supervised manner. (Sun et al. 2021) incorporates the low-loss sample selection strategy with label distribution learning to capture the information in high loss instances. The above co-training based methods focus on single-label noisy learning, which is not suitable to multi-label data.

The Proposed Method

Given ML-PL dataset $\mathcal{D} = \{(\mathcal{X}, \mathcal{Y})\}$, where $\mathcal{X} = \{\boldsymbol{x}_i|_{i=1}^n\}$ are n training instances and $\mathcal{Y} = \{\boldsymbol{y}_j|_{j=1}^n\}$ are their observed labels. For each instance $\boldsymbol{x}_i$, the observed labels $\boldsymbol{y}_i = [y_{i,1}, y_{i,2}, ..., y_{i,C}]^T \in [0,1]^C$ is a C-dimensional vector with C classes, where $y_{i,j} = 1$ indicates instance $\boldsymbol{x}_i$ is annotated with label j, and $y_{i,j} = 0$ otherwise. Given a mini-batch data $\hat{\mathcal{D}} = \{(\hat{\mathcal{X}}, \hat{\mathcal{Y}})\}$, we divide $\hat{\mathcal{Y}} = \{\hat{\boldsymbol{y}}_1, \hat{\boldsymbol{y}}_2, ..., \hat{\boldsymbol{y}}_{\mathcal{B}}\}$ into a positive label set $\hat{\mathcal{Y}}^P = \{\hat{\boldsymbol{y}}_1^P, \hat{\boldsymbol{y}}_2^P, ..., \hat{\boldsymbol{y}}_{\mathcal{B}}^P\}$ and a negative label set $\hat{\mathcal{Y}}^N = \{\hat{\boldsymbol{y}}_1^N, \hat{\boldsymbol{y}}_2^N, ..., \hat{\boldsymbol{y}}_{\mathcal{B}}^N\}$ satisfying $\hat{\boldsymbol{y}}_i^P \cap \hat{\boldsymbol{y}}_i^N =$ and $\hat{\boldsymbol{y}}_i^P \cup \hat{\boldsymbol{y}}_i^N = \hat{\boldsymbol{y}}_i$. We formulate CLS with two networks denoted by $f(\boldsymbol{x}, \Theta_f)$ and $g(\boldsymbol{x}, \Theta_g)$, where Θ_f and Θ_g are the

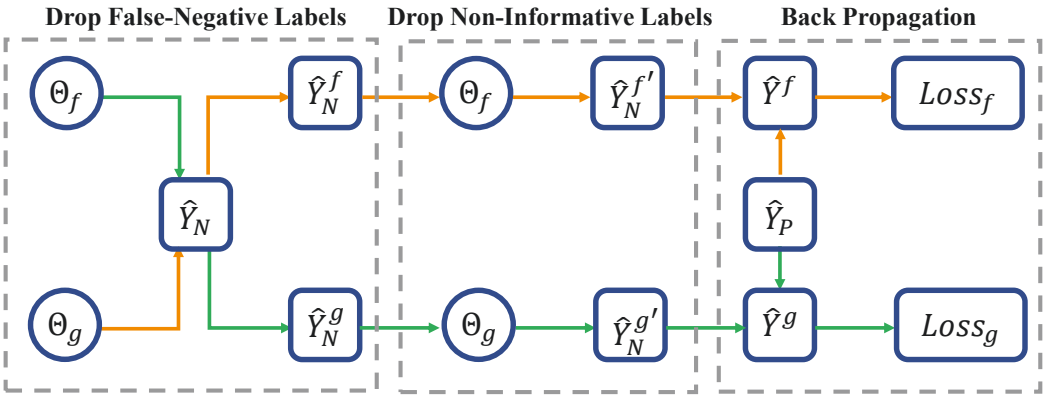


Figure 2: The overall framework of CLS. Given a ML-PL dataset, we let two networks identify false-negative labels for each other to eliminate the negative effect of missing labels and avoid confirmation bias. The well-known positive-negative imbalance issue is also resolved by dropping non-informative negative labels by network itself.

parameters of two networks. Our CLS aims to leverage the two networks to automatically identify false-negative and non-informative labels, accordingly eliminating the negative impact of missing label and avoiding confirmation bias.

Motivation

The widely used *small-loss criterion*, which treats data with minimal loss as clean ones, is justified by the fact that deep neural networks have a tendency to learn simple patterns early before eventually overfitting to noisy samples (Arpit et al. 2017). Thus, if we only use small loss samples in each mini-batch data to train our model, it would be resistant to noisy data. Nonetheless, most samples in ML-PL are annotated with partial labels and tend to own high loss. Simply dropping high loss items would waste numbers of training data and further degenerate the performance of learning model. Based on the above observations, we manipulate the networks to cooperate at label-level rather than sample-level, which just removes non-informative negative labels and reserves training samples to a considerable degree, so as to improve the generalization of the learning model. The overall framework of our proposed CLS is shown in Figure 2.

Algorithm Description

Orthogonal to traditional co-training based methods that two networks cooperate at sample-level and back propagate the selected data from its peer network, we carefully design a new learning manner between two networks, which aims to encourage the two networks to cooperate at label-level and select proper labels for each other, accordingly reserving training data in great extent and improving generalization ability of model. Specifically, we let two peer networks identify true negative labels for each other according to the ranking of label losses within a mini-batch as follows:

$$\hat{\mathcal{Y}}_N^f = \mathrm{Sel}_{min:R_h(e)|\hat{\mathcal{Y}}_N|}\mathcal{L}(\hat{\mathcal{Y}}_N;\hat{\mathcal{P}}^g), \tag{1}$$

$$\hat{\mathcal{Y}}_N^g = \mathrm{Sel}_{min:R_h(e)|\hat{\mathcal{Y}}_N|}\mathcal{L}(\hat{\mathcal{Y}}_N;\hat{\mathcal{P}}^f) \tag{2}$$

where Sel_{min} indicates selecting $R_h(e)|\hat{\mathcal{Y}}_N|$ number of labels with smallest losses from $\hat{\mathcal{Y}}_N$, $R_h(e)$ represents the reserving rate when removing high loss negative labels, e is the round of training epoch, $\mathcal{L}$ is the loss function, $\hat{\mathcal{P}}^g$ and $\hat{\mathcal{P}}^f$ are the prediction probabilities corresponding to $\hat{\mathcal{Y}}_N$ from two networks $f(\boldsymbol{x}, \Theta_f)$ and $g(\boldsymbol{x}, \Theta_g)$, $|\Delta|$ is the number of elements contained in set Δ. By removing high loss negative labels, we reduce the risk of misleading by false-negative labels and mitigate the negative impact of partial annotation.

The inherent positive-negative imbalance in multi-label data has been proved to result in under-emphasizing gradients from positive labels during training stage (Ridnik et al. 2021). To balance the gradients from positive and negative labels, we design a scheme that networks remove non-informative (low loss) negative labels for themselves as:

$$\hat{\mathcal{Y}}_N^{f\prime} = \mathrm{Sel}_{max:R_l(e)|\hat{\mathcal{Y}}_N^f|}\mathcal{L}(\hat{\mathcal{Y}}_N^f;\hat{\mathcal{P}}^{f\prime}), \tag{3}$$

$$\hat{\mathcal{Y}}_N^{g\prime} = \mathrm{Sel}_{max:R_l(e)|\hat{\mathcal{Y}}_N^g|}\mathcal{L}(\hat{\mathcal{Y}}_N^g;\hat{\mathcal{P}}^{g\prime}), \tag{4}$$

where Sel_{max} means selecting $R_l(e)|\hat{\mathcal{Y}}_N^f|$ (or $R_l(e)|\hat{\mathcal{Y}}_N^g|$) number of labels with largest losses from $\hat{\mathcal{Y}}_N^f$ (or $\hat{\mathcal{Y}}_N^g$), $R_l(e)$ indicates the reserving ratio when removing low loss negative labels, $\hat{\mathcal{P}}^{f\prime}$ and $\hat{\mathcal{P}}^{g\prime}$ represent the prediction probabilities corresponding to $\hat{\mathcal{Y}}_N^f$ and $\hat{\mathcal{Y}}_N^g$ from $f(\boldsymbol{x}, \Theta_f)$ and $g(\boldsymbol{x}, \Theta_g)$ respectively. With a proper threshold, the gradients from positive labels and negative labels can be balanced, which encourages the model to treat the positive and negative labels equally and improves the robustness of learning model. Meanwhile, removing low loss negative labels can also avoid model being over-confident on negative labels, which further decreases the risk of model misled by false-negative labels.

In our model, the high loss labels are selected by the peer network to avoid confirmation bias, while the low loss negative labels are selected by the network itself to balance the gradients from positive labels and negative labels. By concatenating the above selected negative labels and all observed positive labels, we obtain the final label sets $\hat{\mathcal{Y}}^f$ =

Algorithm 1: CLS: Co-*Label Selection*

Input: Training Data $\mathcal{D}$; Networks Θ_f and Θ_g; Epoch E_{max} and E_k; Batch Size $\mathcal{B}$; Label Selection Rate τ_h and τ_l, and Learning Rate η;

1: **for** $e = 1, 2, \ldots, E_{max}$ **do**
2: **Shuffle** $\mathcal{D}$ into $\frac{|\mathcal{D}|}{\mathcal{B}}$ mini-batches;
3: **for** i = 1,...,$\frac{|\mathcal{D}|}{\mathcal{B}}$ **do**
4: **Fetch** *i*-th mini-batch $\hat{\mathcal{D}} = \{(\hat{\mathcal{X}}, \hat{\mathcal{Y}})\}$ from $\mathcal{D}$
5: Calculate $\mathcal{P}^f = f(\hat{\mathcal{X}}, \Theta_f)$ and $\mathcal{P}^g = g(\hat{\mathcal{X}}, \Theta_g)$; /* avoid confirmation bias */
6: **Obtain** true negative label sets $\hat{\mathcal{Y}}_N^f$ by (1) in $\hat{\mathcal{Y}}^N$, and $\hat{\mathcal{Y}}_N^g$ by (2) in $\hat{\mathcal{Y}}^N$, respectively; /* resolve positive-negative label imbalance */
7: **Obtain** informative label set $\hat{\mathcal{Y}}_N^{f\prime}$ by (3) in $\hat{\mathcal{Y}}_N^f$, and $\hat{\mathcal{Y}}_N^{g\prime}$ by (4) in $\hat{\mathcal{Y}}_N^f$, respectively;
8: **Obtain** label set $\hat{\mathcal{Y}}^f$ by (5) and $\hat{\mathcal{Y}}^g$ by (6);
9: **Calculate** $\mathcal{L}^f$ by (7) on $\hat{\mathcal{Y}}^f$, and $\mathcal{L}^g$ by (8) on $\hat{\mathcal{Y}}^g$;
10: **Update** $\Theta_f = \Theta_f - \eta\nabla\mathcal{L}^f$ and $\Theta_g = \Theta_g - \eta\nabla\mathcal{L}^g$;
11: **end for**
12: **Update** $R_l(e) = 1 - \min\{\frac{e}{E_k}\tau_l, \tau_l\}$ and $R_h(e) = 1 - \min\{\frac{e}{E_k}\tau_h, \tau_h\}$
13: **end for**

Output: Θ_f and Θ_g

$\{\hat{\boldsymbol{y}}_1^f, \hat{\boldsymbol{y}}_2^f, ..., \hat{\boldsymbol{y}}_{\mathcal{B}}^f\}$ and $\hat{\mathcal{Y}}^g = \{\hat{\boldsymbol{y}}_1^g, \hat{\boldsymbol{y}}_2^g, ..., \hat{\boldsymbol{y}}_{\mathcal{B}}^g\}$ for $f(\boldsymbol{x}, \Theta_f)$ and $g(\boldsymbol{x}, \Theta_g)$ as follows:

$$\hat{\mathcal{Y}}^f = \hat{\mathcal{Y}}_N^{f\prime} \cup \hat{\mathcal{Y}}^P, \tag{5}$$

$$\hat{\mathcal{Y}}^g = \hat{\mathcal{Y}}_N^{g\prime} \cup \hat{\mathcal{Y}}^P. \tag{6}$$

After selecting the final supervised signals for each network, we calculate the loss for $f(\boldsymbol{x}, \Theta_f)$ and $g(\boldsymbol{x}, \Theta_g)$ on these selected labels and their corresponding prediction probabilities $\boldsymbol{p}_i^f$ and $\boldsymbol{p}_i^g$ for further back propagation:

$$\mathcal{L}^f = \frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} \mathcal{L}(\hat{\boldsymbol{y}}_i^f; \boldsymbol{p}_i^f), \tag{7}$$

$$\mathcal{L}^g = \frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} \mathcal{L}(\hat{\boldsymbol{y}}_i^g; \boldsymbol{p}_i^g). \tag{8}$$

Algorithm 1 describes the training process of our proposed CLS, where we train two deep neural networks in a mini-batch manner. It is worth mentioning that we update reserving rate $R_l(e)$ and $R_h(e)$ according to the number of training epochs (Step 12 in Algorithm 1). At the beginning of training, we reserve more training samples (with large reserving rate $R_l(e)$ and $R_h(e)$) in each mini-batch since deep networks would fit clean data first (Wei et al. 2020). With the increasing of epochs, we gradually decrease reserving rate $R_l(e)$ and $R_h(e)$, and maintain fewer training signals in each mini-batch until reaching $1-\tau_l$ and $1-\tau_h$ respectively. Such operation will prevent deep networks from over-fitting noisy data (Han et al. 2018). By cooperating two networks at label-level instead of sample-level and letting them select clean labels for each other, CLS can not only alleviate the negative impact of incomplete annotation, but also reserve training data in most extent, which improves the performance of learning model. Meanwhile, label selection operation in CLS can effectively avoid our model being misled by false-negative labels, especially for the extreme case of ML-PL, where instances are annotated with only one relevant label. Based on the above observations, we argue that the improvement of CLS in comparison with traditional methods should increase more significantly as the missing rate growing, and such argument has clearly been supported by later section.

Strategy	[1]	[2]	[3]	[4]	Ours
Cross Update	✔	✗	✔	✗	✔
Joint Training	✗	✗	✗	✔	✗
Disagreement	✗	✔	✔	✗	✗
Agreement	✗	✗	✗	✔	✗
Label Selection	✗	✗	✗	✗	✔
Sample Selection	✔	✔	✔	✔	✗

Table 1: Comparisons between our proposed CLS and other related approaches. "Cross Update": update parameters in a cross manner instead of parallel manner; "Joint Training": train two networks with a joint loss; "Disagreement": update two networks only on disagreed examples; "Agreement": maximize the agreement of two networks by regularization; "Label Selection": regard high loss labels as false negative labels; "Sample Selection": regard high loss samples as noisy samples. [1]: Co-teaching (Han et al. 2018); [2]: Decoupling (Malach and Shalev-Shwartz 2017); [3]: Co-teaching+ (Yu et al. 2019); [4]: JoCoR (Wei et al. 2020).

Relations to Other Approaches

To further illustrate the connections among co-training approaches, we compare CLS with other related approaches in Table 1. Specifically, Co-teaching (Han et al. 2018) trains two networks in a "Cross-Update" manner to reduce the accumulated error flow while Decoupling (Malach and Shalev-Shwartz 2017) selects instances following the "Disagreement" manner. Co-teaching+ (Yu et al. 2019) incorporates the "Disagreement" strategy and the "Cross Update" strategy to achieve better performance. JoCoR (Wei et al. 2020) selects small loss examples while updates the network parameters by "Joint Training", which contains a co-regularization scheme to maximize the agreement between two networks. Compared with "Disagreement" strategy, Joint Training achieves better performance by adopting an explicit regularization to increase the effective number of selected training data during model co-training. However, it still directly drops possible noisy samples while with meaningful supervised signals, causing the waste of training data. Based on this observation, we convert the cooperation strategy between two networks from "Sample Selection" strategy to "Label Selection", which avoids directly dropping samples and reserves training data to a considerable degree, further improving the performance of learning model.

MS-COCO	75% labels left			50% labels left			25% labels left			single label		
	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1	mAP	CF1	OF1
ML-GCN	71.4	63.1	66.5	67.5	45.5	51.1	63.3	37.6	30.3	64.2	38.6	32.0
CTRAN	72.1	**64.5**	**70.0**	68.8	49.0	52.4	64.4	33.3	29.0	65.3	31.0	29.9
ASL	72.5	62.4	67.7	69.7	49.5	54.0	65.4	34.8	29.1	66.2	31.2	23.7
Co-teaching	72.7	63.0	68.3	69.7	49.3	52.0	65.3	37.2	31.4	65.8	34.8	24.5
Co-teaching+	72.4	64.3	68.6	69.5	52.3	53.6	65.2	39.9	36.1	65.6	35.7	25.6
JoCoR	72.7	63.2	69.0	70.2	50.7	56.9	66.1	33.9	26.7	66.1	31.5	23.4
WAN	67.7	40.3	37.0	65.5	43.5	39.2	62.0	45.2	41.1	63.9	53.3	51.7
ROLE	72.2	59.5	63.4	69.5	46.4	50.2	65.4	34.9	31.3	66.4	38.2	30.6
CLS	**72.8**	64.1	69.3	**70.9**	**65.7**	**70.4**	**67.2**	**54.5**	**60.4**	**67.4**	**61.2**	**66.3**
PASCAL VOC 2007	**75% labels left**			**50% labels left**			**25% labels left**			**single label**		
	mAP	**CF1**	**OF1**	**mAP**	**CF1**	**OF1**	**mAP**	**CF1**	**OF1**	**mAP**	**CF1**	**OF1**
ML-GCN	86.0	78.9	80.2	83.6	72.5	75.3	80.5	68.3	64.3	83.5	67.7	69.5
CTRAN	87.8	80.1	82.2	85.2	75.2	77.1	82.1	70.0	65.1	84.9	69.9	70.2
ASL	88.4	63.2	59.9	86.0	63.3	58.6	82.7	58.7	55.9	85.8	67.5	63.9
Co-teaching	88.4	80.7	**83.7**	86.6	76.3	76.9	83.9	70.8	71.3	86.5	71.5	70.1
Co-teaching+	88.4	80.7	81.5	86.7	77.1	78.1	84.3	71.7	72.0	86.5	73.7	73.0
JoCoR	88.6	80.4	81.8	87.0	77.1	77.7	84.1	70.3	71.4	86.3	71.3	70.7
WAN	85.5	70.2	65.1	83.0	70.1	67.0	79.2	67.4	65.6	84.2	75.2	74.0
ROLE	**89.3**	72.0	71.6	88.0	72.5	72.6	84.5	71.1	71.9	87.7	77.9	79.6
CLS	89.0	**82.0**	83.5	**88.1**	**80.6**	**83.0**	**85.2**	**77.6**	**80.1**	**88.6**	**81.0**	**83.4**
NUS-WIDE	**75% labels left**			**50% labels left**			**25% labels left**			**single label**		
	mAP	**CF1**	**OF1**	**mAP**	**CF1**	**OF1**	**mAP**	**CF1**	**OF1**	**mAP**	**CF1**	**OF1**
ML-GCN	57.3	49.8	64.2	55.0	40.2	50.3	51.2	28.8	32.2	51.5	28.7	30.2
CTRAN	58.1	51.2	67.5	56.1	41.5	52.0	51.6	30.0	33.8	52.1	30.0	31.8
ASL	58.7	**56.1**	67.4	56.6	**57.2**	**69.5**	52.7	**54.2**	**68.7**	53.2	**56.0**	**69.0**
Co-teaching	57.8	50.8	66.8	55.8	41.0	51.0	51.5	29.8	33.1	52.0	30.2	31.0
Co-teaching+	58.0	50.5	67.0	56.0	41.2	51.8	51.5	30.0	32.2	51.9	26.8	31.2
JoCoR	58.3	51.1	67.9	56.2	41.8	52.2	52.1	29.9	34.8	52.4	28.1	33.2
WAN	55.3	38.0	44.2	53.7	40.9	49.9	50.0	37.2	46.2	50.1	38.0	46.5
ROLE	57.7	43.8	53.7	55.2	43.7	54.6	50.5	42.2	55.2	50.9	48.0	63.2
CLS	**58.9**	52.1	**70.7**	**57.0**	49.3	69.1	**53.4**	45.2	66.6	**54.1**	48.0	67.2

Table 2: Comparisons with state-of-the-art methods on MS-COCO, PASCAL VOC 2007 and NUS-WIDE datasets. The best results are presented in bold and the second-best are in underline.

Experiments

Experiment Setup

Datasets We employ three multi-label datasets, including **MS-COCO** (Lin et al. 2014), **PASCAL VOC 2007** (Everingham et al. 2010) and **NUS-WIDE** (Chua et al. 2009). Since these datasets are fully annotated, we follow (Chen et al. 2021b) to randomly drop some positive labels to generate ML-PL datasets according to a dropping rate α, where $\alpha \in \{25\%, 50\%, 75\%\}$ indicates the proportion of dropped positive labels. Besides, we consider the extreme case of ML-PL, where each instance is annotated with only one relevant label. Following (Cole et al. 2021), we randomly select one positive label for each instance to generate such dataset.

Evaluation Metrics We employ three popular multi-label metrics to evaluate each comparing method, including mean Average Precision (mAP), Overall F1-score (OF1) and per-Class F1-score (CF1) (Shen et al. 2021; Chen et al. 2019).

Implementation Details For fair comparison, we adopt Resnet-50 (He et al. 2016) network pre-trained on ImageNet (Deng et al. 2009) as feature extraction backbone for all methods. The input images are squished and randomly cropped into 224×224. Adam is used as the optimizer with a weight decay of 10^{-4}. The batch size is set to 120 for all datasets. We run 100 epochs in total with an initial learning rate of 4×10^{-5} and decrease it to 0 using cosine decay. We adopt Binary Cross Entropy loss as our loss function. We set the ratio of selection rate $R_h(e)$ and $R_l(e)$ as follows: $R_h(e)$ = 1 - min$\{\frac{e}{E_k}\tau_h, \tau_h\}$, $R_l(e)$ = 1 - min$\{\frac{e}{E_k}\tau_l, \tau_l\}$, where τ_l = 0.02 for PASCAL VOC 2007 dataset, and τ_l = 0.06 for NUS-WIDE dataset and MS-COCO dataset. E_k is set to 10 for all datasets. The values of τ_h are {0.002, 0.003, 0.005, 0.005}, {0.005, 0.010, 0.012, 0.012}, and {0.006, 0.012, 0.020, 0.020} on PASCAL VOC 2007 dataset, MS-COCO dataset, and NUS-WIDE dataset under four configurations respectively, which are found through cross-validation. Our

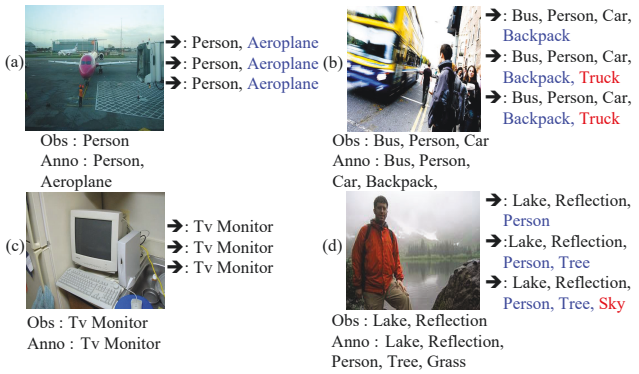


Figure 3: The qualitative results of our CLS.

method is implemented based on PyTorch. We train our model in an end-to-end manner and accomplish all experiments on a computer with an Intel (R) Xeon (R) CPU E5-2620, 64 GB main memory, and two TITAN Xp GPUs.

Comparison with State-of-the-Art Methods

We employ eights methods from three categories for comparison: 1) **ML-GCN** (Chen et al. 2019), **CTRAN** (Lanchantin et al. 2021) and **ASL** (Ridnik et al. 2021), which achieve state-of-the-art performances on classical multi-label image recognition task. 2) **Co-teaching** (Han et al. 2018), **Co-teaching+** (Yu et al. 2019) and **JoCoR** (Wei et al. 2020), which adopt "Sample Selection" strategy to handle noisy label learning. We replace their original loss function with Binary Cross Entropy loss to satisfy ML-PL problem. 3) **WAN** (Cole et al. 2021) and **ROLE** (Cole et al. 2021), which achieve state-of-the-art performances on solving the extreme case of ML-PL, where each instance is annotated with only single positive label. Table 2 reports the experimental results on all employed datasets. We can observe:

- On **MS-COCO** dataset, CLS performs the best in all four cases. Although few methods can achieve comparable performance under the configuration of 75% labels left, CLS significantly outperforms them in other three cases. Especially, as the missing rate increases, the improvements become more significant, e.g. mAP improves 1.1% and 1.3% when training with 25% labels left and single label in comparison with JoCoR.
- On **PASCAL VOC 2007** dataset, CLS outperforms most comparing methods under all configurations, e.g., it outperforms advancing ASL and ML-GCN by 2.8% and 5.1% under the configuration of single label. Although CLS is 0.3% inferior to ROLE under the configuration of 75% labels left, we attribute such phenomenon to the property of PASCAL VOC 2007 that it just covers 5011 training images from 20 categories, which is more simple than MS-COCO and NUS-WIDE. For other configuaration cases, CLS outperforms ROLE by 0.1%, 0.7%, and 0.9% respectively, which demonstrates that our proposed method is robust to false-negative labels.
- On **NUS-WIDE** dataset, CLS obtains the best performance over all four cases. Especially, for such heavy positive-negative label imbalance data, CLS outperform JoCoR by 0.6%, 0.8%, 1.3%, 1.7%, respectively. Meanwhile, it also outperforms state-of-the-art single positive methods ROLE by 3.2% under the case of single positive label configuration. Besides, in comparison with ASL method, which introduces a novel asymmetric loss to handle the positive-negative imbalance, CLS still outperforms it by 0.2%, 0.4%, 0.7%, 0.9% under four cases. These results strongly verify the effectiveness of our proposed method on solving label imbalance problem.

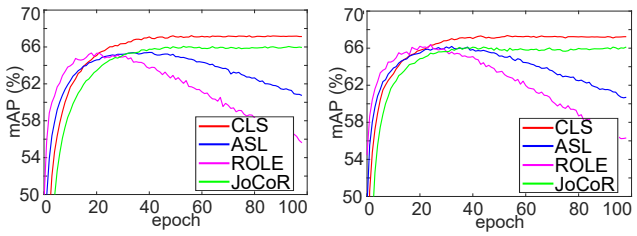


Figure 4: Results of visualization on MS-COCO dataset.(a) 25% labels left; (b) single label.

Further Analysis

Qualitative Results We present the qualitative results of CLS in Figure 3. *Obs* means the observed positive labels in training set, *arrow* indicates the positive labels predicted by our CLS during training, and *Anno* indicates annotated positive labels in the original dataset. The blue labels indicate the corresponding labels are identified as false negative labels and belong to the annotated labels, while the red labels indicate the corresponding labels are identified as false negative labels but do not belong to the annotated labels. It can be seen that although only a few ground truth positive labels are available, our method can still distinguish most false-negative and true-negative labels. Specifically, our CLS successfully identifies four false-negative labels in total five absent annotated positive labels. Although our CLS confuses the similar category *Car* and *Truck* in Figure 3.(b), we observe that CLS can correctly identify *Sky* as false-negative labels, which is not annotated in the original dataset. Such phenomenon verifies that CLS successfully keeps the model from memorizing false-negative labels.

Visualization on eliminating confirmation bias We show mAP vs. epochs on MS-COCO in Figure 4. In all four plots, we can observe the memorization effect of networks. Specifically, the performance of single-network based methods (ASL and ROLE) increases rapidly in the first few epochs, but decreases significantly in the following epochs. In contrast, co-training based methods (CLS and JoCoR) can well alleviate or even prevent the decreasing trend, which demonstrates the superiority of co-training strategy on eliminating confirmation bias. Meanwhile, for co-training based JoCoR, our proposed CLS outperforms it in all four cases. We attribute such success to "Label Selection" cooperation strategy that can exploit most training data in comparison with "Sample Selection" strategy.

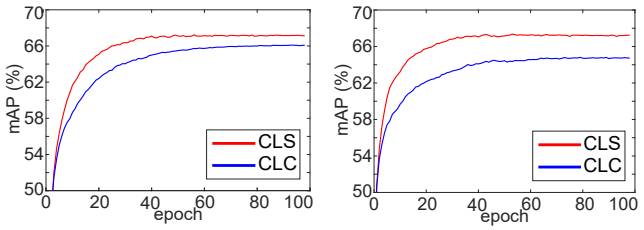


Figure 5: Comparison with Co-Label Correction on MS-COCO. (a) 25% labels left; (b) single label.

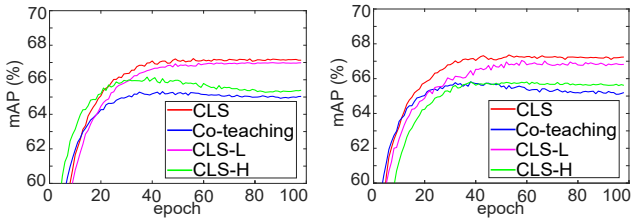


Figure 6: Results of ablation studies on MS-COCO dataset. (a) 25% labels left; (b) single label.

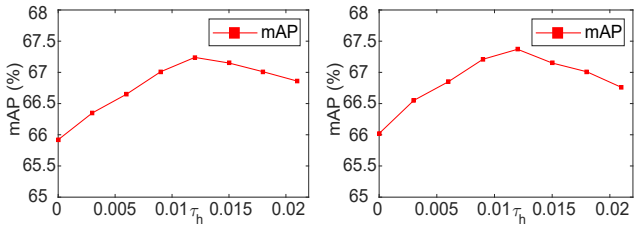


Figure 7: Accuracy comparisons with different τ_h on MS-COCO. (a) 25% labels left; (b) single label.

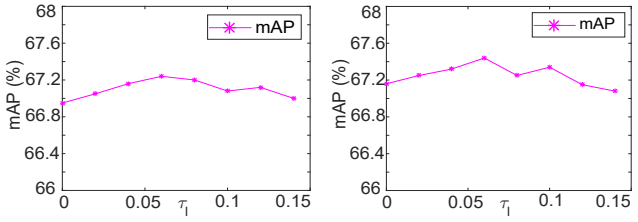


Figure 8: Accuracy comparisons with different τ_l on MS-COCO. (a) 25% labels left; (b) single label.

Label Correction or Label Selection? In our model, we let the two networks remove false-negative labels for subsequent model training. A straightforward question arises that **what if we correct false-negative labels**. To verify this idea, we conduct an experiment on MS-COCO dataset. As shown in Figure 5, although Co-*Label Correction* (CLC) reaches comparable performance with CLS in the first 5 epochs, its performance is significantly inferior to CLS in all the subsequent epoches. Such phenomenon come from the label correction operation that CLC corrects labels according to the ranking of label loss within a mini-batch, which introduces misleading supervised information (see Figure 3.(c) for an example), further degrading model performance. However, CLS just removes non-informative negative labels, which can effectively alleviate the negative influence of missing label and improve the generalization of learning model.

Ablation Studies We conduct ablation study to evaluate the impact of key components in CLS. Specifically, "CLS-H" means our CLS without the module of high loss negative labels removing and "CLS-L" means our CLS without the module of low loss negative labels removing. As illustrated in Figure 6, we find "CLS-H" suffers significant performance drop in comparison with CLS, which verifies the negative impact of false-negative labels can be alleviated by removing high loss negative labels. We also find that CLS consistently outperforms "CLS-L" during the entire training epochs, which demonstrates that removing non-informative negative labels alleviates positive-negative imbalance and forces the model to focus on informative labels. Meanwhile, the phenomenon that "Co-teaching" performs worst in last 50 epochs evaluates the superiority of "Label Selection" strategy in comparison with "Sample Selection".

Hyperparameters Sensitivity Analysis To explore the effects of different values of hyperparameters, we conduct sensitivity analysis for τ_l and τ_h on MS-COCO dataset. We vary the estimated noise rate τ_h among {0, 0.003, 0.006, ..., 0.021}, and show the results in Figure 7. As shown in Figure 7, when filtering out the negative labels of high losses (i.e. false negative labels), multi-label recognition accuracy is boosted. However, when too many negative labels are filtered out, the accuracy drops since high loss true negative labels are also ignored. Similarly, we vary the values of the dropping rate of low loss negative labels τ_l among {0, 0.02, 0.04, ..., 0.14}. As shown in Figure 8, balancing the weights between positive labels and negative labels by dropping low loss negative labels has positive contribution to model performance, where it achieves the best balance when $\tau_l = 0.06$. Note that filtering out high loss true negative labels will degrade model performance since model have not learned from these labels enough, while filtering out low loss true negative labels can balance the weights between positive and negative labels, and force model to learn more from informative labels, further improving model performance.

Conclusion

In this paper, we proposed an effective approach termed CLS to tackle ML-PL problem, which trains two networks simultaneously and let them remove false-negative labels for each other to alleviate the negative effect of incomplete annotation and avoid confirmation bias. Since "small sample loss" criteria would drop more valuable trainig data in ML-PL, we convert the correlation strategy between peer networks from "Sample Selection" to "Label Selection", accordingly reserving training data in most extent. By removing non-informative negative labels, the highly imbalance between positive and negative labels is also alleviated. Extensive results have verified the effectiveness of CLS.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2023YFB3107100), the National Natural Science Foundation of China (No. 62306020), the Young Elite Scientist Sponsorship Program by BAST (BYESS2024199), and the Beijing Natural Science Foundation (No. 4242046, L244009), and the Major Research Plan of National Natural Science Foundation of China (No. 92167102).

References

Arpit, D.; Jastrzębski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A.; and Bengio, Y. 2017. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, 233–242.

Chen, T.; Pu, T.; Wu, H.; Xie, Y.; and Lin, L. 2021a. Structured Semantic Transfer for Multi-Label Recognition with Partial Labels. *arXiv preprint arXiv:2112.10941*.

Chen, Z.; Wei, X.; Wang, P.; and Guo, Y. 2019. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5177–5186.

Chen, Z.; Wei, X.; Wang, P.; and Guo, Y. 2021b. Learning Graph Convolutional Networks for Multi-Label Recognition and Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 6969–6983.

Chua, T.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. Nus-wide: a real-world web image database from national university of singapore. In *ACM International Conference on Image and Video Retrieval*, 1–9.

Cole, E.; Mac Aodha, O.; Lorieul, T.; Perona, P.; Morris, D.; and Jojic, N. 2021. Multi-Label Learning from Single Positive Labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 933–942.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Durand, T.; Mehrasa, N.; and Mori, G. 2019. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 647–657.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 303–338.

Feng, L.; An, B.; and He, S. 2019. Collaboration based multi-label learning. In *Proceedings of the AAAI conference on artificial intelligence*, 3550–3557.

Ge, Z.; Guan, Y.; Li, X.; and Fu, B. 2022. Consistent, Balanced, and Overlapping Label Trees for Extreme Multi-label Learning. In *Proceedings of the ACM International Conference on Information & Knowledge Management*, 551–560.

Gu, Z.; Feng, S.; Hu, R.; and Lyu, G. 2023. ONION: Joint Unsupervised Feature Selection and Robust Subspace Extraction for Graph-based Multi-View Clustering. *ACM Transactions on Knowledge Discovery from Data*, 17(5): 1–23.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Proceedings of the Advances in Neural Information Processing Systems*, 1–9.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Huynh, D.; and Elhamifar, E. 2020. Interactive multi-label CNN learning with partial labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9423–9432.

Jia, B.; and Zhang, M. 2023. Multi-dimensional multi-label classification: Towards encompassing heterogeneous label spaces and multi-label annotations. *Pattern Recognition*, 138: 109357.

Lanchantin, J.; Wang, T.; Ordonez, V.; and Qi, Y. 2021. General Multi-label Image Classification with Transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 16478–16488.

Li, J.; Socher, R.; and Hoi, S. 2020. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *Proceedings of the International Conference on Learning Representations*, 1–9.

Lin, T.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 740–755.

Liu, B.; Jia, B.; and Zhang, M. 2023. Towards enabling binary decomposition for partial multi-label learning. *IEEE transactions on pattern analysis and machine intelligence*, 45(11): 13203–13217.

Liu, W.; Wang, H.; Shen, X.; and Tsang, I. 2021. The emerging trends of multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (11): 7955–7974.

Liu, W.; Yuan, J.; Lyu, G.; and Feng, S. 2023. Label driven latent subspace learning for multi-view multi-label classification. *Applied Intelligence*, 53(4): 3850–3863.

Lyu, G.; Kang, W.; Wang, H.; Li, Z.; Yang, Z.; and Feng, S. 2024a. Common-Individual Semantic Fusion for Multi-View Multi-Label Learning. In *International Joint Conference on Artificial Intelligence*, 1–9.

Lyu, G.; Yang, Z.; Deng, X.; and Feng, S. 2024b. L-VSM: Label-Driven View-Specific Fusion for Multiview Multilabel Classification. *IEEE Transactions on Neural Networks and Learning Systems*, 1–15.

Malach, E.; and Shalev-Shwartz, S. 2017. Decoupling" when to update" from" how to update". In *Proceedings of the Advances in Neural Information Processing Systems*, 961–971.

Min, C.; Lin, H.; Li, X.; Zhao, H.; Lu, J.; Yang, L.; and Xu, B. 2023. Finding hate speech with auxiliary emotion detection from self-training multi-label learning perspective. *Information Fusion*, 96(214): 214–223.

Ridnik, T.; Ben-Baruch, E.; Zamir, N.; Noy, A.; Friedman, I.; Protter, M.; and Zelnik-Manor, L. 2021. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE International Conference on Computer Vision*, 82–91.

Shen, X.; Dong, G.; Zheng, Y.; Lan, L.; Tsang, I. W.; and Sun, Q. 2021. Deep co-image-label hashing for multi-label image retrieval. *IEEE Transactions on Multimedia*, 24: 1116–1126.

Sun, Y.; Zhang, Y.; and Zhou, Z. 2010. Multi-label learning with weak label. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 593–598.

Sun, Z.; Liu, H.; Wang, Q.; Zhou, T.; Wu, Q.; and Tang, Z. 2021. Co-LDL: A Co-Training-Based Label Distribution Learning Method for Tackling Label Noise. *IEEE Transactions on Multimedia*, 1093–1104.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the Advances in Neural Information Processing Systems*, 1195–1204.

Vu, X.; Le, D.; Edlund, C.; Jiang, L.; and Nguyen, H. D. 2020. Privacy-Preserving Visual Content Tagging using Graph Transformer Networks. In *Proceedings of the ACM International Conference on Multimedia*, 2299–2307.

Wang, H.; Peng, C.; Dong, H.; Feng, L.; Liu, W.; Hu, T.; Chen, K.; and Chen, G. 2024. On the Value of Head Labels in Multi-Label Text Classification. *ACM Transactions on Knowledge Discovery from Data*, 18(5): 1–21.

Wang, J.; Feng, S.; Lyu, G.; and Gu, Z. 2023. Triple-Granularity Contrastive Learning for Deep Multi-View Subspace Clustering. In *ACM International Conference on Multimedia*, 2994–3002.

Wang, S.; Wang, J.; Wang, Z.; and Ji, Q. 2015. Multiple emotion tagging for multimedia data by exploiting high-order dependencies among emotions. *IEEE Transactions on Multimedia*, 2185–2197.

Wang, Z.; Fang, Z.; Li, D.; Yang, H.; and Du, W. 2022. Semantic Supplementary Network With Prior Information for Multi-Label Image Classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4): 1848–1859.

Wei, H.; Feng, L.; Chen, X.; and An, B. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 13726–13735.

Wu, X.; Jiang, B.; Zhong, Y.; and Chen, H. 2022. Multi-target Markov boundary discovery: Theory, algorithm, and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4964–4980.

Yu, X.; Han, B.; Yao, J.; Niu, G.; Tsang, I.; and Sugiyama, M. 2019. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, 7164–7173.

Zhong, Q.; Lyu, G.; and Yang, Z. 2024. Align While Fusion: A Generalized Nonaligned Multiview Multilabel Classification Method. *IEEE Transactions on Neural Networks and Learning Systems*, 1–10.