Trusted Loss Correction for Noisy Multi-Label Learning

Amirmasoud Ghiassi* S.GHIASSI@TUDELFT.NL
Delft University of Technology, Delft, Netherlands

Cosmin Octavian Pene* PENE.COSMIN.OCTAVIAN@GMAIL.COM
Delft University of Technology, Delft, Netherlands

Robert Birke ROBERT.BIRKE@UNITO.IT
Computer Science dept., University of Turin, Turin, Italy

Lydia Y.Chen LYDIAYCHEN@IEEE.ORG
Delft University of Technology, Delft, Netherlands

Editors: Emtiyaz Khan and Mehmet Gönen

Abstract

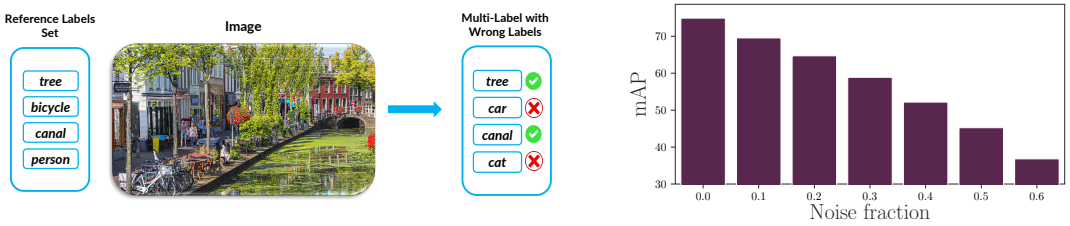
Noisy and corrupted labels are shown to significantly undermine the performance of multi-label learning, which has multiple labels in each image. Correcting the loss via a label corruption matrix is effective in improving the robustness of single-label classification against noisy labels. However, estimating the corruption matrix for multi-label problems is no mean feat due to the unbalanced distributions of labels and the presence of multiple objects that may be mapped into the same labels. In this paper, we propose a robust multi-label classifier against label noise, TLCM, which corrects the loss based on a corruption matrix estimated on trusted data. To overcome the challenge of unbalanced label distribution and multi-object mapping, we use trusted single-label data as regulators to correct the multi-label corruption matrix. Empirical evaluation on real-world vision and object detection datasets, i.e., MS-COCO, NUS-WIDE, and MIRFLICKR, shows that our method under medium (30%) and high (60%) corruption levels outperforms state-of-the-art multi-label classifier (ASL) and noise-resilient multi-label classifier (MPVAE), by on average 12.5% and 26.3% mean average precision (mAP) points, respectively.

Keywords: Multi Label Learning; Corrupted Labels; Corruption Matrix Estimation; Deep Neural Network

1. Introduction

Real-world images, collected from different resources, e.g., social media and web search engines, naturally contain multiple objects corresponding to diverse labels which can be annotated by crowdsourcing. While the crowd offers an inexpensive and scalable solution to curate datasets to power up deep neural networks (DNN), it also suffers from notorious label noise issues which may result in severe degradation on the learning performance (Hendrycks et al., 2018; Chen et al., 2019a). Multi-label learning, where images have multiple objects and complex object-label mappings, further exacerbate the noisy label problem (Zhao and Gomes, 2021). For instance, (Veit et al., 2017) shows that the Open Images dataset (Krasin

*. These authors contributed equally to this work



(*a*) Example of image with wrong labels

(*b*) Impact of increasing wrong label ratios

Figure 1: Wrong labels and their impact in multi-label classification.

et al., 2016), which is widely used for multi-label and multi-class image classification, contains 26.6% false positive labels among the training set.

Label noise and its impact on the performance of DNNs have been widely studied in single-label classification (Algan and Ulusoy, 2021; Hendrycks et al., 2018; Yao et al., 2019). It is shown that DNNs can overfit to noise degrading their performance significantly (Zhang et al., 2017) due to the memorization effect (Xu et al., 2021). In multi-label learning, where each sample may include several corrupted labels as shown in Figure 1(*a*), label noise becomes a more consequential obstacle to learn accurate classifiers (Zhao and Gomes, 2021). We empirically evaluate the effect of label corruption by injecting different noise levels on the MS-COCO dataset[1] used to train a state-of-the-art multi-label classifier, i.e., ASL (Ridnik et al., 2021b). In Figure 1(*b*), one can clearly see that the mean Average Precision (mAP) significantly degrades with increasing noise levels.

Most prior work focuses on learning classifiers robust to label noise in single-label problems using different techniques such as filtering out the noisy samples (Han et al., 2018a; Yu et al., 2019), and correcting the loss via an estimated corruption matrix (Hendrycks et al., 2018; Patrini et al., 2017b). Although these robust methods are effective in single-label learning, their performances against noisy labels in multi-label problems are limited. To verify their effectiveness, we choose two different robust methods including filtering noisy labels (co-teaching (Han et al., 2018a)) and estimating noise patterns (GLC (Hendrycks et al., 2018)) via trusted data. We extend co-teaching and GLC to robust multi-label classification and use them to train a classifier on the MS-COCO dataset with 40% label noise achieving an mAP of 50.34 and 54.88, respectively. Only GLC barely improves the achieved mAP compared to non noise-resilient training method. Filtering based methods, e.g., co-teaching, use informativeness measures for ranking the samples, then select clean candidates for training. However, the single-label informativeness measures fail in most cases (Wu et al., 2020). The adapted GLC provides a slight noise resilience for the multi-label classification method by correcting the loss using a corruption matrix estimation. However, the estimated matrix does not capture the corruption probability accurately. The reasons are two folds. First, the single-label classification assumption is based on the independence of images and corresponding labels. In contrast, there is an arbitrary number of objects and labels and possible correlations among them in multi-label classification. Second, the frequency of labels across

1. The details of noise injection process can be found in Section 4.1

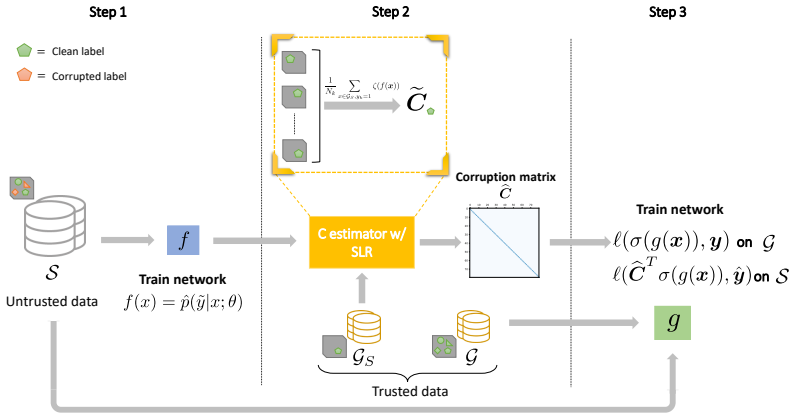


Figure 2: Overview of TLCM

samples in multi-label datasets makes it hard to estimate an accurate corruption probability for each label.

To cope with label corruption in multi-label learning, we propose a novel approach which uses gold loss correction leveraging a small fraction of trusted (gold) data for estimating the label corruption matrix. Although the gold loss correction has been used effectively against single-label corruption (Hendrycks et al., 2018), noisy multi-label data needs an accurate method to estimate the corruption matrix handling the label mapping complexity and the label unbalance difficulty. We overcome these challenges by taking advantage of trusted single-labels to regulate the estimation of the multi-label corruption matrix. In other words, we introduce single-label regulators to correct the captured matrix on the miss classification behavior of the trained model on noisy multi-label samples. It is worth mentioning, having a small amount of trusted data, i.e. clean single-label and multi-label data, is common practice because we use clean label data in the validation and testing process. Therefore, it is possible to curate a small amount of training data. To the best of our knowledge, our method is the first that estimates the label corruption matrix in multi-label classification considering the labels dependence and unbalance.

To such an end, we design a loss correction algorithm via a novel corruption matrix estimation with a single-label regulator for multi-label classification called TLCM. First, we estimate the label corruption matrix pre-training a network on untrusted data and using it on trusted multi-label samples. Next, we correct our estimated corruption matrix via regulating the corruption probabilities with the corruption probabilities calculated using trusted single-label samples. We exploit single-label regulators to eliminate the faulty effect of unbalanced labels and lack of conditional label independence. Ultimately, we correct the prediction of a multi-label classifier via the corruption matrix. We empirically demonstrate the performance of TLCM on a number of multi-labels datasets, e.g., MS-COCO (Lin et al., 2014), NUS-WIDE (Chua et al., 2009) and MIRFLICKR (Huiskes and Lew, 2008) with various level of label corruption, against ASL (Ridnik et al., 2021b) and MPVAE (Bai et al., 2020). TLCM can achieve remarkable mAP compared to the state-of-the-art, just by using 10% of trusted data, TLCM improves mAP by average 30%, 20.25% and 18.39% on

under extremely high noise ratio, i.e., 60%, for MS-COCO, NUS-WIDE, and MIRFLICKR, respectively. The contributions of this paper are summarized as follows:

- We design a label corruption estimation technique that uses trusted multi-label and single-label data in order to calculate the corruption matrix considering label dependence and unbalance.
- We design a robust multi-label classifier, TLCM, built on a loss correction approach using our label corruption matrix estimation with single-label regulators.
- We compare TLCM against state-of-the-art classifiers, i.e., ASL and MPVAE, under noisy labels and study the behavior of TLCM in target ablation study experiments.

1.1. Motivation example

Our motivation stems from the detrimental effects of label noise in training data can have on the model performance. We demonstrate this using the state-of-the-art ASL (Ridnik et al., 2021b) method to train a TResNet-M (Ridnik et al., 2021c) network on the MS-COCO dataset (Lin et al., 2014). ASL applied on TResNet ranks top on the leader board for multi-label classification on MS-COCO[2]. We inject symmetric label noise (details in Section 4.1) at various corruption levels, from 0% to 60%, and report the mean average precision to assess the impact of wrong labels. mAP is considered by many recent works (Lanchantin et al., 2020; Chen et al., 2019b) as an important metric for performance evaluation in multi-label classification since it takes into consideration both false-negative and false-positive rates (Ridnik et al., 2021b). Figure 1(*b*) shows the results: each additional 10% noisy labels leads to a 5%-8% reduction in mAP score. Since it is hard and costly to avoid label noise (Zhao and Gomes, 2021), it is vital to develop robust classifiers that can avoid overfitting to the label noise in the training data.

2. Related Work

Recent literature has shown increased interest towards robustness against noisy labels in training, much more in single-label classification than in multi-label learning. We first investigate the robust learning solutions in single-label classification, followed by an analysis of the multi-label context.

2.1. Noisy Single-label Classification

Several prior works tackle the problem of noisy labels in a single-label classification where each sample has single corresponding labels with a probability of corruption. (Song et al., 2020) distinguishes between five categories of robust DNNs. Some classifiers such as C-model (Goldberger and Ben-Reuven, 2017), Contrastive-Additive Noise Network (Yao et al., 2019) and Robust Generative Classifier (RoG) (Lee et al., 2019) rely on a noise adaptation layer at the top of the softmax layer. Another solution is to use regularization techniques such as data augmentation (Shorten and Khoshgoftaar, 2019), and batch normalization (Ioffe and Szegedy, 2015). These methods perform well on low to moderate noise, but

2. https://paperswithcode.com/sota/multi-label-classification-on-ms-coco visited June 24, 2021.

fail on datasets with higher noise (Tanno et al., 2019). Other works use sample selection showing impressive results even under heavy noise. Examples are MentorNet (Jiang et al., 2018), Co-teaching (Han et al., 2018a) and Co-teaching+ (Yu et al., 2019). Another stream of work estimates the noise corruption matrix to correct the labels during training, without, e.g. Forward (Patrini et al., 2017a), LABELNET (Ghiassi et al., 2021a), and Masking (Han et al., 2018b), or with, corrected loss e.g. TrustNet (Ghiassi et al., 2021b), and Golden Loss Correction (Hendrycks et al., 2018), a small portion of trusted, i.e. verified clean, data. All the mentioned methods can only perform well with single-label classification tasks and they are not noise-tolerant models for multi-label learning.

2.2. Noisy Multi-label Classification

In multi-label learning, little attention has been given to the consequences of label noise (Song et al., 2020; Xie and Huang, 2022). Few papers treat noisy and missing labels in the multi-label context (Zhang et al., 2021). (Sun et al., 2019) learns by a low-rank and sparse decomposition approach to obtain ground-truth and irrelevant label matrices. (Fang and Zhang, 2019) introduces label confidence to recover the true labels. (Xie and Huang, 2020) formalizes the two objectives of recovering ground-truth and identifying noisy labels via a unified regulators-based framework. The topic is gaining traction with some more works available as technical reports. (Zhao and Gomes, 2021) leverages context to identify noisy labels. (Li et al., 2020) proposes a two-step noise correction. (Bai et al., 2020) acknowledges and evaluates the impact of noisy labels even if it is not the main goal of the authors.

In contrast to the above, TLCM aims to leverage single-label regulators together with a small fraction of trusted data to avoid overfitting to noisy labels in multi-label classification.

3. Methodology

Consider the multi-label dataset $\mathcal{D}$ comprising N tuples $(\boldsymbol{x}, \tilde{\boldsymbol{y}})$ where $\boldsymbol{x} \subset \mathbb{R}^d$ denotes one sample with d features and $\tilde{\boldsymbol{y}} = [\tilde{y}_1, ..., \tilde{y}_k, ..., \tilde{y}_K]$ denotes the corresponding label vector over K classes with binary elements $\tilde{y}_k \in [0, 1]$. The label vectors are affected by noise, hence a $\tilde{y}_k$ can be the true (y_k) or a corrupted $(\bar{y}_k)$ label. We assume that a subset of the data, i.e., the *gold* dataset $\mathcal{G} \subset \mathcal{D}$, can be trusted having no (or for practical purposes negligible) corrupted labels. We refer to the rest of the data with potentially corrupted labels as *silver* dataset $\mathcal{S} = \mathcal{D} - \mathcal{G}$. We define the *trusted fraction* as the ratio $\eta = \frac{|\mathcal{G}|}{|\mathcal{G}|+|\mathcal{S}|}$. Furthermore, we assume that a small dataset of clean single-label images $\mathcal{G}_S$ is available. We use $\mathcal{G}$ and $\mathcal{G}_S$ to estimate a $K \times K$ noise corruption matrix $\widehat{\boldsymbol{C}}$ characterizing the label noise. Each element $\hat{c}_{ij}$ of noise corruption matrix $\widehat{\boldsymbol{C}}$ represents the probability of a label of class i to be flipped into a label of class j, formally:

$$\hat{c}_{ij} = p(\tilde{y}_j = 1 \wedge \tilde{y}_i = 0 | y_j = 0 \wedge y_i = 1)$$

Finally, similar to related work, e.g. (Ridnik et al., 2021b), we treat the multi-label classification problem as a series of binary classification tasks, one for each label class.

3.1. Overview of TLCM

We propose a novel *Trusted Loss Correction for Noisy Multi-Label Learning* (TLCM) approach for noise resilient multi-label training. To address the challenge of noisy multi-labels, TLCM uses a gold loss correction which uses a small set of trusted (gold) samples to estimate the noise corruption matrix used to correct the model predictions during training. The gold loss correction has been shown to be highly effective against single-label noise. However, multi-label noise exacerbates the difficulty to estimate the noise corruption matrix. The fact that each sample can have an arbitrary number of labels leads to two main challenges. First, single-label classification assumes conditional independence of $\boldsymbol{y}$ given $\boldsymbol{x}$. In the single-label setting this assumption holds since $\boldsymbol{y}$ is deterministic in $\boldsymbol{x}$. Due to having an arbitrary number of labels and possible correlation between labels, this assumption does not hold in the multi-label setting which makes it hard to separate out which label or labels lead to a specific corrupted label. Second, the frequency distribution of labels across samples is harder to control. This leads to intrinsically higher label unbalances in multi-label datasets. For example, the popular and commonly used single-label CIFAR datasets have equal numbers of samples for each class. On the contrary, the ratio between the most (person) and least (toaster) frequent label in the MS-COCO dataset is 1197.4. This unbalance influences the noise. The more common a label, the higher the impact of that label on the noise characteristics.

Both the lack of conditional label independence and the label unbalance lead to poor estimates of multi-label noise corruption matrices. To overcome this and estimate the true multi-label noise corruption matrix, TLCM introduces the use of single-label regulators. These capture the miss classification behavior of a model trained on noisy data for specific labels and are used to correct the estimation of the multi-label noise corruption matrix.

Figure 2 presents an overview of the TLCM method for noise resilient multi-label classification. It includes three main steps. Step 1: we train a *silver* classifier $f(\cdot;\theta)$ on the noisy samples in $\mathcal{S}$. Step 2, the heart of TLCM, consists of two substeps. First, we capture the single-label noise characteristics in $\widetilde{\boldsymbol{C}}$ using f on the trusted single-label samples in $\mathcal{G}_S$. Second, we use f on the trusted samples in $\mathcal{G}$ to compute a multi-label noise corruption matrix $\widehat{\boldsymbol{C}}$ regularized by $\widetilde{\boldsymbol{C}}$ to counter the detrimental effect of multi labels on noise estimation. Step 3: we train the *gold* classifier $g(\cdot;\phi)$ on the samples from $\mathcal{S}$ with label predictions corrected via the noise corruption matrix $\widehat{\boldsymbol{C}}$, plus the samples from $\mathcal{G}$, to maximize the impact of the trusted data.

3.2. Noise Corruption Matrix Estimation

***Silver* classifier**. Before starting the estimation of the multi-label corruption matrix, we need to train a *silver* classifier $f(\boldsymbol{x};\theta) = \widehat{p}(\tilde{\boldsymbol{y}}|\boldsymbol{x})$ on $\mathcal{S}$. Given the labels in $\mathcal{S}$ are potentially corrupted, f is not a reliable classifier for our final predictions. However, we can use f for our multi-label noise corruption matrix estimation.

It is worth mentioning that our estimation method for multi-label noise corruption matrix is independent of the loss and network types used to train $f(\cdot;\theta)$. Here, we consider asymmetric loss (Ridnik et al., 2021b) due to its superior performance to counter the detrimental effect of irrelevant class labels in multi-label datasets. For each class (we omit the

class index k for brevity) the loss is:

$$\mathcal{L}_{ASL} = \begin{cases} L_+ = (1-p)^{\gamma_+}\log(p), & y_k = 1 \\ L_- = (p_m)^{\gamma_-}\log(1-p_m), & y_k = 0 \end{cases}$$

where L_+ and L_- are the positive and negative loss parts used for relevant and irrelevant labels, respectively, p is the network output probability and γ_+, γ_- are the focusing parameters. The focusing parameters control the weights of positive and negative labels. Finally, $p_m = \max(p - m, 0)$ denotes the shifted probability where the margin m is a tunable hyper-parameter controlling the contribution of irrelevant labels introduced by (Ridnik et al., 2021b).

Noise corruption matrix. Once we trained the *silver* classifier $f(\cdot;\theta)$ on $\mathcal{S}$ we can start the multi-label noise corruption matrix estimation. First, we use f on the samples in $\mathcal{G}_S$ to capture the single-label noise characteristics via the single-label noise corruption matrix $\widetilde{\boldsymbol{C}}$. For each label $k \in K$ the corresponding row of $\widetilde{\boldsymbol{C}}$ is given by the averaged softmax output of f for all samples $\boldsymbol{x} \in \mathcal{G}_S$ with $y_k = 1$:

$$\widetilde{\boldsymbol{C}}_{k\cdot} = \frac{1}{N_k} \sum_{x \in \mathcal{G}_S, y_k = 1} \zeta(f(\boldsymbol{x}))$$

where $\widetilde{\boldsymbol{C}} \subset \mathbb{R}^{K \times K}$, $\widetilde{\boldsymbol{C}}_{k\cdot}$ denotes the k^{th} matrix row, N_k the number of samples with $y_k = 1$, and $\zeta(\cdot)$ the softmax function. We scale the output via softmax to enhance the training since here we only consider single-label samples.

Next we use $\widetilde{\boldsymbol{C}}$ to regulate the label correlations from the multi-label samples in $\mathcal{G}$. To do this we compute the multi-label noise corruption matrix using f on the samples in $\mathcal{G}$ while regulating each time the output of f via $\widetilde{\boldsymbol{C}}$. For each label $k \in K$ the corresponding row of $\widehat{\boldsymbol{C}}$ is given by the averaged sigmoid output of f for all samples $\boldsymbol{x} \in \mathcal{G}$ having $y_k = 1$ regulated by the difference in the single-label corruption probabilities of label k and the other labels present in $\boldsymbol{x}$:

$$\widehat{\boldsymbol{C}}_{k\cdot} = \frac{1}{N_k} \sum_{(\boldsymbol{x},\boldsymbol{y}) \in \mathcal{G}, y_k = 1} \left(\sigma(f(\boldsymbol{x})) + \sum_{\forall l \neq k, y_l = 1} \left(\widetilde{\boldsymbol{C}}_{k\cdot} - \widetilde{\boldsymbol{C}}_{l\cdot} \right) \right) \tag{1}$$

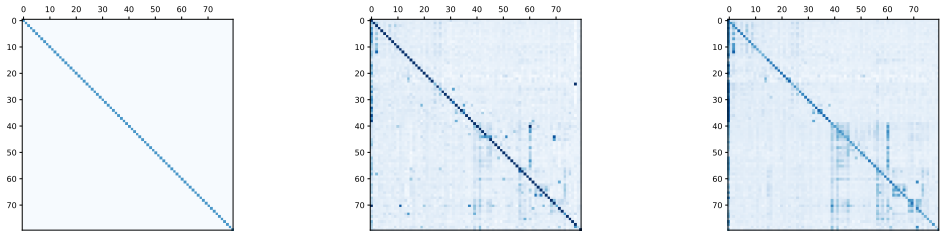
where $\sigma(\cdot)$ denotes the sigmoid function.

Figure 3 compares the multi-label corruption matrix estimated by TLCM for 40% symmetric label noise against the one injected –ground truth–, and the one estimated without regularization. The latter corresponds to equation (1) without the internal summation term for rows of $\widetilde{\boldsymbol{C}}$. The comparison highlights the benefits of single-label regularization. This can be seen from the darker, closer to the truth, diagonal values and the more pronounced difference with respect to the off-diagonal values.

***Gold* classifier**. With the estimated multi-label noise corruption matrix $\widehat{\boldsymbol{C}}$, we finally train the robust *gold* classifier $g(\cdot;\phi)$. We correct labels of the samples in $\mathcal{S}$ via $\widehat{\boldsymbol{C}}$ while leveraging samples in $\mathcal{G}$ as is. The loss function follows as:

$$\begin{aligned} \ell &= \mathcal{L}_{ASL}(\widehat{\boldsymbol{C}}^T \sigma(g(\boldsymbol{x})), \hat{\boldsymbol{y}}), \quad \forall x \in \mathcal{S} \\ \ell &= \mathcal{L}_{ASL}(\sigma(g(\boldsymbol{x})), \boldsymbol{y}) \qquad \forall x \in \mathcal{G}. \end{aligned}$$

More detail about our novel noise estimation method and our robust model (TLCM) can be found in the supplementary material.



(a) True (b) TLCM (c) GLC

Figure 3: Comparison between multi-label corruption matrices with 40% noise on MS-COCO.

4. Evaluation

4.1. Experiment setup

Datasets. We evaluate TLCM using three popular multi-label vision and object detection datasets, i.e. MS-COCO (Lin et al., 2014), NUS-WIDE (Chua et al., 2009) and MIR-FLICKR (Huiskes and Lew, 2008).

- **MS-COCO** is a popular real-world dataset of common objects in context widely used for evaluation of multi-label classification. The training and validation datasets contain 82K and 40K images, respectively. Each image is tagged on average with 2.9 labels belonging to 80 classes.
- **NUS-WIDE** is a web image dataset with 269.6K manually annotated samples in 81 visual concepts. In our experiments, since some URLs have been eliminated, we were able to download 218K images and split into 186K as training and 32K testing samples. Images have 2.9 labels on average.
- **MIRFLICKR** is an open image collection with $25K$ images annotated in 38 categories. The training and testing sets include 23,300 and 1,700 images, respectively. Here, the average labels per image is 8.9.

To explicitly test TLCM in the more challenging multi-label setting, we remove all single label images from both the training and validation datasets. This does not only elicit a more reliable evaluation, but it also allows to collect single-label samples to construct $\mathcal{G}_S$. After filtering the number of images per class label varies from 1, for unpopular classes, to 1,234, for the most popular class with an average of 210.2 images per class for MS-COCO and from 4 to 40,026 and from 112 to 9,613 for NUS-WIDE and MIRFLICKR, respectively. More single-label samples allow to estimate more accurate regulators which in turn leads to a more robust classifier. Finally, we split the training data into a *gold* ($\mathcal{G}$) and a *silver* ($\mathcal{S}$) datasets. As the base, we use 10% as gold data, leading to 6.5K clean samples and 58.7K samples injected with noisy labels for MS-COCO. Similarly we have 7.5K clean and 67.5K noisy samples for NUS-WIDE and 2K clean and 18K noisy samples for MIRFLICKR.

Label Noise. Label noise in multi-label data is more complex than in a single-label context since each sample has an arbitrary number of labels. We follow previous works (Jiang

et al., 2018; Patrini et al., 2017a) and inject symmetric noise, but with an extra step. Specifically, we select a fraction η, i.e. the noise ratio, of labels and flip them to another class with uniform probability. This corresponds to a noise corruption matrix $\boldsymbol{C}$ having elements c_{ij} as follows:

$$c_{ij} = \begin{cases} 1-\eta & \text{if } i = j \\ \dfrac{\eta}{K-1} & \text{if } i \neq j \end{cases}$$

In order to ensure wrong label injection, we test whether or not the new label is already associated with the image. If it does, we repeatedly elect a new label until we select one which is not yet present. In order to evaluate how robust TLCM is to noise, we test our method against multiple noise ratios –from 0% to 60%.

Evaluation Metrics. For a comprehensive and reliable evaluation, we follow conventional settings and report the following metrics: mean average precision (mAP), average per-class F1 (CF1) and average overall F1 (OF1). These metrics have been widely used in literature to evaluate multi-label classification (Ridnik et al., 2021b; Song et al., 2020; Wang et al., 2020) and have been shown to dramatically decrease with label noise (Zhao and Gomes, 2021). Note that only the training set is affected by noise, whereas the evaluation metrics are computed on the clean testing set.

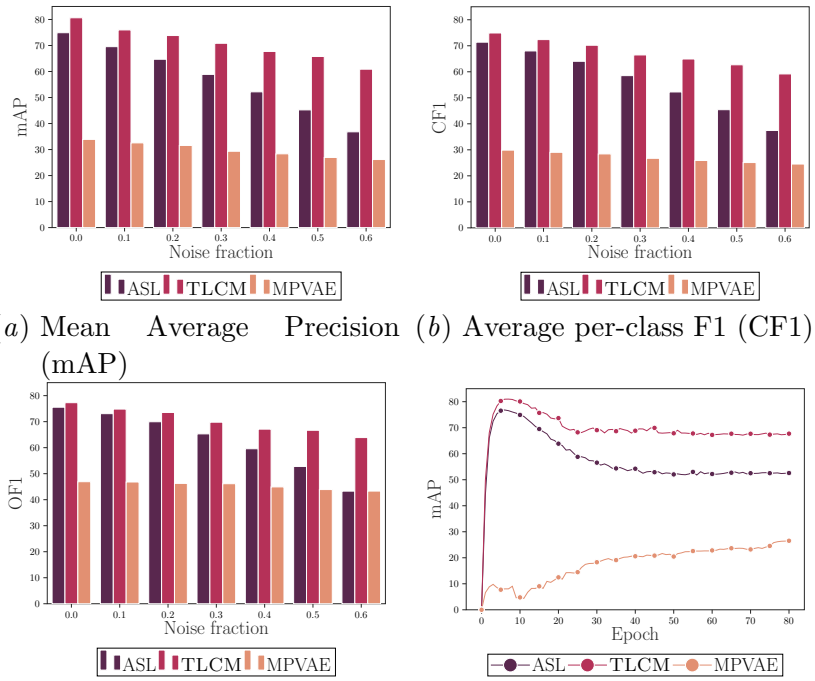
DNN Architecture. As base architecture for the DNN, we use TResNet (Ridnik et al., 2021c). TResNet network is a high performance GPU-dedicated architecture based on ResNet50 designed to increase the model prediction performance without increasing training or inference time. The TResNet network is pre-trained on the ImageNet-21K dataset for better generalizability and increased prediction accuracy (Ridnik et al., 2021a). In particular we use the TResNet-M version with input resolution 224 for MS-COCO and MIRFLICKR datasets. We use the TResNet-L version with input resolution 448 for NUS-WIDE dataset. The encoder and decoder use the structure from (Bai et al., 2020), i.e, 3-layer fully connected neural networks with ReLU activation function. Moreover, we set the hyper-parameters to the default values provided in (Bai et al., 2020) for each dataset.

Baseline. We compare the performance of TLCM against MPVAE (Bai et al., 2020) which is a noise resilient multi-label model, and ASL (Ridnik et al., 2021b) which is one of the state-of-the-art multi-label learning algorithms.

- **MPVAE** (Bai et al., 2020) proposes a variational autoencoder based method that encodes the features and labels to Gaussian subspaces and then decodes the samples from the subspaces into multivariate probit to predict the image labels.
- **ASL** (Ridnik et al., 2021b) introduces new asymmetric loss for multi-label classification to reduce the impact of irrelevant labels in focal loss (Lin et al., 2020) and balance the weights of relevant and irrelevant labels.

For a fair comparison, we test ASL, MPVAE and TLCM on the same datasets with the same label noise. TLCM assumes access to a small subset of clean samples. Thus the only additional knowledge of our method is which labels are trusted, i.e., belonging to the small golden dataset $\mathcal{G}$, and which labels are potentially corrupted.

Implementation Details. We use PyTorch v1.9.0 for all the methods, and the default parameters provided in (Ridnik et al., 2021b; Bai et al., 2020) except that we always take



(*a*) Mean Average Precision (mAP)

(*b*) Average per-class F1 (CF1)

(*c*) Average overall F1 (OF1)

(*d*) Memorization effect (40% noise)

Figure 4: Evaluation of TLCM, ASL and MPVAE on MS-COCO with symmetric label noise using 10% of trusted data.

the last trained model due to the *memorization effect*. The number of training epochs is an important parameter for a reliable evaluation, especially in a noisy setting. DNNs are shown to present the so-called *memorization effect* (Xu et al., 2021; Feldman and Zhang, 2020; Feldman, 2020) benefiting in general from this factor to achieve a better prediction performance in atypical samples. However, (Arpit et al., 2017) suggests that with noisy data, DNNs prioritize learning simple patterns first. From preliminary experiments we see that 80 epochs are enough for the learning to stabilize.

4.2. Results

In this subsection we empirically compare the performance of TLCM to two rivals, i.e., ASL and MPVAE under 0% to 60% symmetric noise. We aim to show the effectiveness of our TLCM in robustly learning from noisy data.

Figure 4 shows the comparison results. The performance of all systems decreases under increasing noise levels, but TLCM is significantly more robust. In terms of mAP TLCM consistently outperforms ASL and MPVAE for all noise ratios (see Figure 4(*a*)). After TLCM the best mAP performance is achieved by ASL followed finally by MPVAE, under all noise levels. Without noise, i.e. 0% noise ratio, the mAP of TLCM and ASL are extremely close, but MPVAE performance is still significantly lower. Since MPVAE uses a Variational Auto Encoder based method and the network architecture is less deep and

complex than ASL and TLCM, the MPVAE can not perform well. It is worth mentioning that the network architecture is taken from (Bai et al., 2020). ASL's performance drops an average of 5.34% points with each 10% noise, while TLCM's performance decreases by only 3.29% points. Also the mAP degradation for MPVAE is 1.28% points. Under severe noise, i.e. 60%, the gap between TLCM and ASL is more than 24% points and only 14.01% points worse than without noise. In comparison ASL drops by 38.1% points from 0% to 60% noise. This shows that TLCM is robust even to high noise levels. Similar results apply for both CF1 and OF1, see Figure 4(b) and Figure 4(c), respectively. Even if ASL is slightly close to TLCM in the no noise case, the performance quickly degrades with additional noise. At 60% noise TLCM is better than ASL by 21.73% and 20.59% points for CF1 and OF1, respectively. Moreover, the CF1 and OF1 for MPVAE slightly degrade from 29.0% and 46.8% to 24.5% and 43.3%, respectively, over increasing noise ratios.

To reliably assess the correctness of TLCM, we also investigate the observed memorization effect for TLCM (depicted in Figure 4(d)). Both TLCM and ASL follow the same trend. First they learn the easy patterns, achieving a high accuracy after just a few epochs. However, afterwards the performance slowly degrades over training effort and finally stabilizes after approximately 60 epochs. But the mAP trend for MPVAE is different. It slightly increases over the training epoch and then reaches a steady-state which however is significantly lower than TLCM and ASL. The figure clearly shows the advantage of TLCM over ASL and MPVAE in the different levels at which they plateau. Moreover, one can observe that TLCM has a slight delay in learning at the beginning of the training, i.e. TLCM peaks at epoch 10, while ASL at epoch 6. This observation indicates that TLCM does not help in terms of learning speed nor in reaching a higher performance during training, but by preventing overfit to the noisy labels. This makes the DNN more resistant to wrong label information. Furthermore, this suggests that our method can also be applied to other existing classifiers and domains.

Table 1 summarizes the mAP of TLCM, ASL and MPVAE for the NUS-WIDE and MIRFLICKR datasets under 0%, 30% and 60% label noise. As illustrated in Table 1, the mAP of TLCM, ASL and MPVAE for NUS-WIDE under all noise ratios are significantly lower than for MS-COCO (as well as MIRFLICKR, see Table 1). NUS-WIDE includes images with bigger sizes, one additional class and higher label imbalance. This makes NUS-WIDE a more complex dataset for multi-label classification. TLCM is still able to provide robustness for training against label noise and beats both baselines. ASL and TLCM perform similarly at 0% noise but ASL degrades faster under increasing noise. MPVAE is less affected by increasing noise but plateaus at a significantly lower level. By using only 10% trusted samples, TLCM obtains 38.7% mAP under severe label corruption, i.e., $\eta = 60\%$, while ASL and MPVAE only achieve 22.6% and 14.2% mAP, respectively. Across the different noise ratios TLCM achieves on average 7% and 16% points higher mAP compared to ASL and MPVAE, respectively. Overall we conclude that while NUS-WIDE is a more challenging dataset the trends are similar to MS-COCO.

Furthermore, Table 1 analyzes the performance of TLCM compared to ASL and MPVAE on MIRFLICKR. This dataset is less challenging than MS-COCO and NUS-WIDE because it only uses about half, i.e. 38, the number of label classes. As a consequence all three multi-label classification methods achieve higher mAP compared to two other datasets. Again TLCM is robust to noise and the best of the three methods achieving on

Table 1: mAP(%) of noisy NUS-WIDE and MIRFLICKR corrupted with 30% and 60% noise for different noise resilient networks

Method	*Percent Trusted*	*NUS-WIDE*			*MIR-FLICKR*		
		$\eta = 0\%$	$\eta = 30\%$	$\eta = 60\%$	$\eta = 0\%$	$\eta = 30\%$	$\eta = 60\%$
ASL	-	60.15	43.59	22.64	80.85	63.06	35.26
MPVAE	-	17.55	15.67	14.21	41.01	39.03	35.68
TLCM	5	59.91	46.62	34.11	80.90	65.72	46.37
	10	60.23	48.56	38.68	80.88	67.34	53.86
	15	60.66	48.79	38.89	80.91	68.83	54.05

average across all noise levels 7.6% and 28.8% points higher mAP than ASL and MPVAE, respectively, using 10% trusted samples. While the general trends are similar to the previous two datasets, interestingly, for $\eta = 60\%$ MPVAE performs slightly better than ASL. This emphasizes that MPVAE is less sensitive to noisy labels. Still MPVAE's mAP is 18.2% points lower than TLCM.

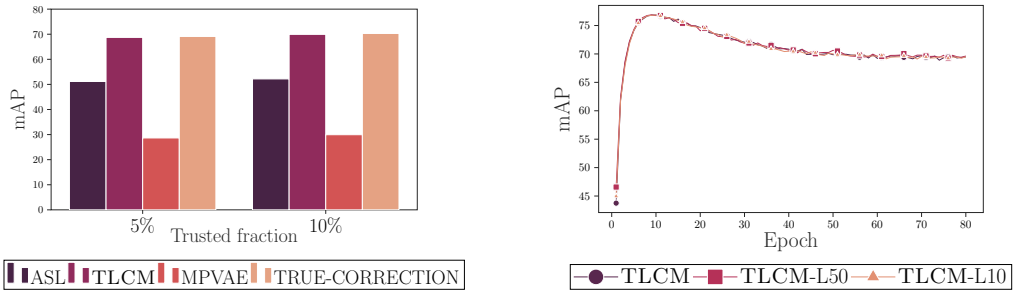
4.3. Ablation Study

To better understand the performance of TLCM, we perform extra ablation studies to investigate the effects of: i) errors in the noise corruption matrix estimation; ii) impact of the gold dataset size (both studied in experiment I); and iii) impact of number of single-label images (studied in experiment II). The base setup of the experiments is the same as in Section 4.1 with only the changes specifically mentioned.

Experiment I: Figure 3 shows visually the difference between the true and our estimated noise corruption matrix. To assess also quantitatively how good our estimation method works, we train classifier g with the true corruption matrix. Figure 5(a) compares the achieved mAP results on MS-COCO under 40% noise. g trained with the true corruption matrix represents the upper performance bound achievable by noise corruption matrix estimators. Since TLCM uses trusted data to estimate the noise corruption matrix and train the robust classifier g, we expect the size of $\mathcal{G}$ to have an impact on the estimation accuracy and consequently on model performance. We investigate this effect by repeating the previous experiments with halve the fraction of trusted data, i.e. 5%. This corresponds to 3,263 clean samples and 62,005 samples injected with noisy labels. Figure 5(a) shows these results via the two different bar plot groups.

Estimating the noise corruption matrix with TLCM, we reach 68.7% and 69.9% mAP under 5% and 10% of trusted data, respectively. Using the true noise correction matrix increases these numbers to 69.1% and 70.3%. The difference between TLCM and the upper bound (True-Correction) is below 0.5% points in both cases. This shows that our noise estimation can capture almost perfectly the impact of the noise corruption matrix, and that it works even with a reduced amount of trusted data. The difference of 1.2% points when using the true noise correction matrix stems from including $\mathcal{G}$ to train the gold classifier $g(\cdot, \phi)$.

Similar trends hold also in the other two datasets (see Table 1). Increasing the size of the trusted data achieves higher mAP scores. This gain increases with the amount of label noise but decreases with the amount of trusted data. For 30% label noise, moving from 5%



(*a*) Impact of trusted data (*b*) Impact of single-label data

Figure 5: Ablation study of TLCM on MS-COCO

trusted data to 10% increases the mAP for NUS-WIDE/MIRFLICKR by 1.9%/1.6% points, but from 10% to 15% only by 0.2%/1.5% points. Instead for 60% label noise, the respective gains are 4.6%/7.5% points and 0.2%/1.1% points. Also gains are generally higher for MIRFLICKR than for NUS-WIDE. MIRFLICKR has fewer classes which eases the noise corruption matrix estimation. Even on different datasets, TLCM is able to provide good performance gains even with little trusted data.

Experiment II: To assess the impact of trusted single-label images on the estimation of the corruption matrix, we conduct two extra experiments with varying images per class for MS-COCO. In addition to the previous case using all single-label images, we limit the number of single-label images per class to at most 50 and 10, referred to as TLCM-L50 and TLCM-L10, respectively. This results in a total of 2824/721 single-label images used for TLCM-L50/TLCM-L10. Figure 5(*b*) shows the impact on the mAP over training epochs. One can observe that limiting the number of single-label images has only a minor impact on the performance of TLCM. Hence our proposed method is not only robust to wrong labels in multi label learning but it also can estimate an accurate noise corruption matrix by using only a small proportion of trusted single-label data. In other words, TLCM has a limited dependency on the amount of clean single-label data.

5. Conclusion

Noisy labels significantly undermine the performance of classification models. While noisy single-label problems are well addressed in the prior art, little is known about the noisy multi-label problems which have unprecedented challenges - highly unbalanced label distributions and complex mapping between objects and labels. In this paper, we propose TLCM that enhances the robustness of multi-label learning against noisy multi-labels by assigning trusted loss correction based on estimated noise corruption matrix. To tackle the unbalanced label distribution without the ground truth of object-label mapping in multi-label problems, we design a novel estimation scheme for noise corruption matrix using a small fraction of the trusted label set. Specifically, we use a subset of single-label images in the trusted set to construct a regulator to rebalance the estimated corruption matrix toward the true values. We evaluate TLCM on three real world datasets subject to different levels of label noise. Under 30% and 60% label corruption TLCM shows an improvement

of mean average precision compared to ASL and MPVAE ranging between 2.7-36.7% and 10.7-38.2% points, respectively.

References

G. Algan and I. Ulusoy. Image classification with deep learning in the presence of noisy labels: A survey. *Knowl. Based Syst.*, 215:106771, 2021.

Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, and et al. A closer look at memorization in deep networks. In *ICML*, volume 70, pages 233–242, 2017.

Junwen Bai, Shufeng Kong, and Carla P. Gomes. Disentangled variational autoencoder based multi-label classification with covariance-aware multivariate probit model. In *IJCAI*, pages 4313–4321, 2020.

Pengfei Chen, Benben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In *ICML*, volume 97, pages 1062–1070, 2019a.

Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Y. Guo. Multi-label image recognition with graph convolutional networks. *CVPR*, pages 5172–5181, 2019b.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS-WIDE: a real-world web image database from national university of singapore. In Stéphane Marchand-Maillet and Yiannis Kompatsiaris, editors, *CIVR*, 2009.

Jun-Peng Fang and Min-Ling Zhang. Partial multi-label learning via credible label elicitation. In *AAAI*, pages 3518–3525. AAAI Press, 2019.

V. Feldman. Does learning require memorization? a short tale about a long tail. *ACM SIGACT Symposium on Theory of Computing*, 2020.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. In *NIPS*, 2020.

Amirmasoud Ghiassi, Robert Birke, Rui Han, and Lydia Y Chen. Labelnet: Recovering noisy labels. In *(IJCNN)*, pages 1–8. IEEE, 2021a.

Amirmasoud Ghiassi, Robert Birke, and Lydia Y.Chen. Trustnet: Learning from trusted data against (a)symmetric label noise. In *(BDCAT '21)*, page 52–62, 2021b.

J. Goldberger and E. Ben-Reuven. Training deep neural-networks using a noise adaptation layer. In *ICLR*, 2017.

B. Han, Quanming Yao, Xingrui Yu, Gang Niu, M. Xu, Weihua Hu, I. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NIPS*, 2018a.

Bo Han, Jiangchao Yao, Gang Niu, Ivor W. Tsang, Ya Zhang, and Masashi Sugiyama. Masking: A new perspective of noisy supervision. In *NIPS*, pages 5841–5851, 2018b.

Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *NIPS*, pages 10456–10465, 2018.

Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *ICMIR*, pages 39–43, 2008.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, volume 37, pages 448–456, 2015.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, pages 2309–2318, 2018.

Ivan Krasin, Tom Duerig, Neil Alldrin, Andreas Veit, Sami Abu-El-Haija, Serge Belongie, David Cai, Zheyun Feng, Vittorio Ferrari, Victor Gomes, Abhinav Gupta, Dhyanesh Narayanan, Chen Sun, Gal Chechik, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. 2016.

Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. *ArXiv*, abs/2011.14027, 2020.

Kimin Lee, Sukmin Yun, Kibok Lee, Honglak Lee, B. Li, and Jinwoo Shin. Robust inference via generative classifiers for handling noisy labels. In *ICML*, 2019.

Junnan Li, Caiming Xiong, Richard Socher, and Steven C. H. Hoi. Towards noise-resistant object detection with noisy annotations. *CoRR*, abs/2003.01285, 2020.

Tsung-Yi Lin, M. Maire, Serge J. Belongie, James Hays, P. Perona, D. Ramanan, Piotr Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):318–327, 2020.

Giorgio Patrini, A. Rozza, A. Menon, R. Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. *CVPR*, pages 2233–2241, 2017a.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, pages 1944–1952, 2017b.

Tal Ridnik, Emanuel Ben Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *CoRR*, abs/2104.10972, 2021a.

Tal Ridnik, Emanuel Ben Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *ICCV*, pages 82–91, 2021b.

Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben Baruch, Gilad Sharir, and Itamar Friedman. Tresnet: High performance gpu-dedicated architecture. In *WACV*, pages 1400–1409, 2021c.

Connor Shorten and Taghi Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6, 07 2019.

Hwanjun Song, Minseok Kim, Dongmin Park, and J. Lee. Learning from noisy labels with deep neural networks: A survey. *ArXiv*, abs/2007.08199, 2020.

Lijuan Sun, Songhe Feng, Tao Wang, Congyan Lang, and Yi Jin. Partial multi-label learning by low-rank and sparse decomposition. In *AAAI*, volume 33, pages 5016–5023, 2019.

Ryutaro Tanno, A. Saeedi, S. Sankaranarayanan, D. Alexander, and N. Silberman. Learning from noisy labels by regularized estimation of annotator confusion. *CVPR*, 2019.

Andreas Veit, N. Alldrin, Gal Chechik, Ivan Krasin, A. Gupta, and Serge J. Belongie. Learning from noisy large-scale datasets with minimal supervision. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. Multi-label classification with label graph superimposing. In *AAAI*, 2020.

Jian Wu, Victor S Sheng, Jing Zhang, Hua Li, Tetiana Dadakova, Christine Leon Swisher, Zhiming Cui, and Pengpeng Zhao. Multi-label active learning algorithms for image classification: Overview and future promise. *CSUR*, 53(2):1–35, 2020.

Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning with noisy label identification. In *AAAI*, pages 6454–6461, 2020.

Ming-Kun Xie and Sheng-Jun Huang. CCMN: A general framework for learning with class-conditional multi-label noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

Han Xu, Xiaorui Liu, Wentao Wang, Wenbiao Ding, Zhongqin Wu, Zitao Liu, Anil Jain, and Jiliang Tang. Towards the memorization effect of neural networks in adversarial training, 2021.

J. Yao, J. Wang, I. Tsang, Ya Zhang, J. Sun, C. Zhang, and R. Zhang. Deep learning from noisy image labels with quality embedding. *IEEE Transactions on Image Processing*, 28: 1909–1922, 2019.

Xingrui Yu, B. Han, Jiangchao Yao, Gang Niu, I. Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *ICML*, volume 97, pages 7164–7173, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.

Youcai Zhang, Yuhao Cheng, Xinyu Huang, Fei Wen, Rui Feng, Yaqian Li, and Yandong Guo. Simple and robust loss design for multi-label learning with missing labels. *arXiv preprint arXiv:2112.07368*, 2021.

W. Zhao and Carla P. Gomes. Evaluating multi-label classifiers with noisy labels. *ArXiv*, abs/2102.08427, 2021.