# Multi-Label Gold Asymmetric Loss Correction with Single-Label Regulators

**Cosmin Octavian Pene**[1] , **Amirmasoud Ghiassi**[1] , **Taraneh Younesian**[1] , **Lydia Y. Chen**[1]
[1]TU Delft

## Abstract

Multi-label learning is an emerging extension of the multi-class classification where an image contains multiple labels. Not only acquiring a clean and fully labeled dataset in multi-label learning is extremely expensive, but also many of the actual labels are corrupted or missing due to the automated or non-expert annotation techniques. Noisy label data decrease the prediction performance drastically. In this paper, we propose a novel Gold Asymmetric Loss Correction with Single-Label Regulators (GALC-SLR) that operates robust against noisy labels. GALC-SLR estimates the noise confusion matrix using single-label samples, then constructs an asymmetric loss correction via estimated confusion matrix to avoid overfitting to the noisy labels. Empirical results show that our method outperforms the state-of-the-art original asymmetric loss multi-label classifier under all corruption levels, showing mean average precision improvement up to 28.67% on a real-world dataset of MS-COCO, yielding a better generalization of the unseen data and increased prediction performance.

## 1 Introduction

Real-world images naturally contain multiple object classes. Multi-label learning is an extension of the multi-class classification where the input image displays multiple labels. It is extremely time-consuming and expensive to collect high-quality labels for single-label images. Even long-standing and highly curated datasets, e.g. CIFAR [17], contain wrong labels [5]. Acquiring a clean fully labeled dataset for multi-label classification is even more challenging. For example, [32] shows that the Open Images dataset [16], which is widely used for multi-label and multi-class image classification, contains 26.6% false positives among the training label set.

In single-label classification, label noise has been widely studied in the literature and its effects have been carefully investigated. [1] suggests that although Deep Neural Networks (DNNs) are somewhat robust to label noise, "their tendency to overfit data makes them vulnerable to memorizing even random noise", resulting in poorer classification performance. There are multiple techniques used for coping with noisy labels in single-label images. However, multi-label classification is a more complex problem. [28] suggests that the simple extensions of existing noise resilient single-label methods are not able to learn the proper correlations among multiple labels.


**Figure 1: Wrong label noise in multi-label classification**

Very little attention has been given to the study of multi-label robust classifiers. We aim to fill this gap in noise-resilient multi-label classifiers by proposing a novel Gold Asymmetric Loss Correction with Single-Label Regulators (GALC-SLR). GALC-SLR assumes that a small subset of the training data can be trusted and uses this additional information to accurately estimate the noise corruption matrix. Due to class imbalance and label correlations, learning the noise in real-world multi-label datasets is more difficult than in real-world single-label datasets. Hence, we introduce a novel method that uses *single-label regulators* to rebalance the predictions towards a targeted label. This leads to accurate noise estimations used to correct the wrong labels during training, making the model robust to label noise even in the most challenging multi-label setting.

Our study specifically studies the effect of wrong labels. As depicted in Fig. 1, each image comes with multi labels including some wrong and some clean ones. This paper aims to answer the following research questions:

- What is the impact of wrong labels on the performance of a state-of-the-art multi-label classifier?
- How to accurately estimate the multi-label noise distribution using extra information from trusted data?
- How to cope with the class imbalance and label correlations, well-known issues in multi-label learning?
- How to train an accurate multi-label classifier with wrong label information?

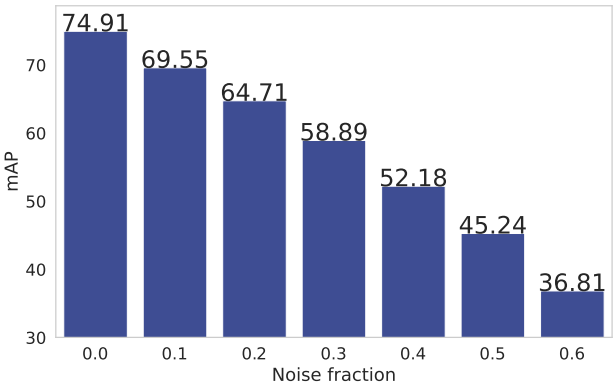



**Figure 2: Impact of increasing wrong label ratios**

In comparison to the state-of-the-art Asymmetric Loss (ASL) multi-label classifier [4] GALC-SLR is significantly more accurate under label noise. ASL balances the probabilities of different samples by treating positive and negative samples differently, i.e. asymmetrically. In empirical evaluation on the MS-COCO dataset [21] GALC-SLR outperforms ASL under all tested noise ratios from 0% to 60%. GALC-SLR improves the mean Average Precision (mAP) over ASL on average by 13.81% and up to 28.67%.

### 1.1 Motivation example

In this section we show to detrimental effects of noisy labels on the performance of a state-of-the-art multi-label classifier, by conducting our own experiment with wrong labels. We demonstrate this by using the ASL [4] method to train a TResNet-M [25] network on the MS-COCO dataset [21]. ASL applied on TResNet ranks top on the leader board for MLC on MS-COCO[1]. 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 [19; 6] an important metric for performance evaluation in multi-label classification since it takes into consideration both false-negative and false-positive rates [4].

Fig. 2 gives a visual representation of the results. As expected, the results show a dramatic decrease in prediction performance, with each additional 10% noise leading to a 5%-8% reduction in mAP score. Hence, it seems that even an accurate state-of-the-art classifier such as ASL suffers from a dramatic decrease in performance when trained on noisy labels. Since it is hard and costly to avoid label noise [38], it is vital to develop robust classifiers that can avoid overfitting 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.

**Single-label Classification** Many researchers have been tackling the problem of noisy labels in single-label classification. [28] provides a comprehensive evaluation of 57 state-of-the-art robust DNNs. The paper distinguishes between five categories of robust DNNs. Some classifiers such as C-model [10], Contrastive-Additive Noise Network [36], and Robust Generative Classifier (RoG) [20] have been shaped to have a Robust Architecture by "adding a noise adaptation layer at the top of the softmax layer". Another solution is to use regularization techniques such as data augmentation [26], weight decay [18], dropout [29], and batch normalization [14]. These methods perform well on low to moderate noise but fail on datasets with higher noise [31]. Other solutions use Sample Selection and have shown impressive results, most of them being robust even to heavy noise. A few popular examples are MentorNet [15] in which a student network relies on a pre-trained mentor network that indicates which labels are likely to be correct, Co-teaching [11] and Co-teaching+ [37] that also use two collaborative DNNs, the latter one introducing decoupling [22], and other hybrid approaches such as SELFIE [27] that combines the sample selection strategy with a loss correction approach. The last category consists of classifiers with Loss Adjustment. These methods are actively modeling the noise distribution by estimating the label corruption matrix and use this information to correct the noisy labels during training. A few popular methods are Forward [23], Masking [12], and Gold Loss Correction (GLC) [13]. Out of the three, GLC uses additional information from trusted data in order to estimate the label noise distribution more accurately.

**Multi-label Classification** In multi-label learning, little attention has been given to the consequences of label noise [28]. Few papers treat noisy labels in the multi-label context. For example, [30] uses a low-rank and sparse decomposition technique to obtain ground-truth and irrelevant label matrices. [7] introduces label confidence to restore the clean labels. [34] uses a unified regulators-based framework to recover the ground-truth labels and to also identify the corrupted ones. Another proposed method leverages context to identify noisy labels [38]. With the problem of noisy labels becoming more and more popular, other papers such as [3] acknowledge the importance of this issue by explicitly conducting experiments with noisy labels. Even though robustness was not the main goal of the authors, they still manage to design a classifier that outperforms other state-of-the-art classifiers even when trained on noisy labels.

In contrast to the methods described above, GALC-SLR 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: GALC-SLR

### 3.1 Notation

Consider the multi-label dataset $\mathcal{D} = \{(\boldsymbol{x}_i, \tilde{\boldsymbol{y}}_i)\}_{i=1}^{N}$ where $\boldsymbol{x}_i \subset \mathbb{R}^d$ denotes the $i^{th}$ sample out of $N$ with $d$ features. $\tilde{\boldsymbol{y}}_i \subset [0,1]^K$ denotes the corresponding label vector over $K$

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

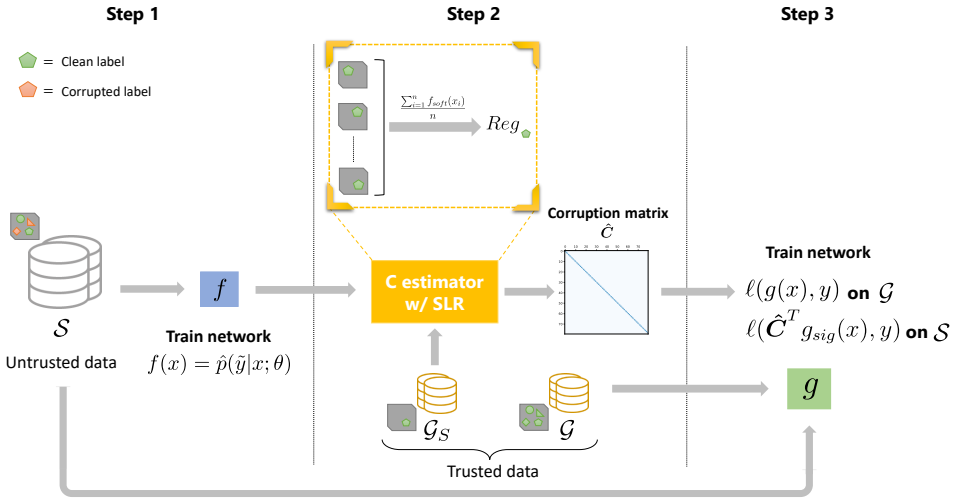



**Figure 3: Overview of GALC-SLR**

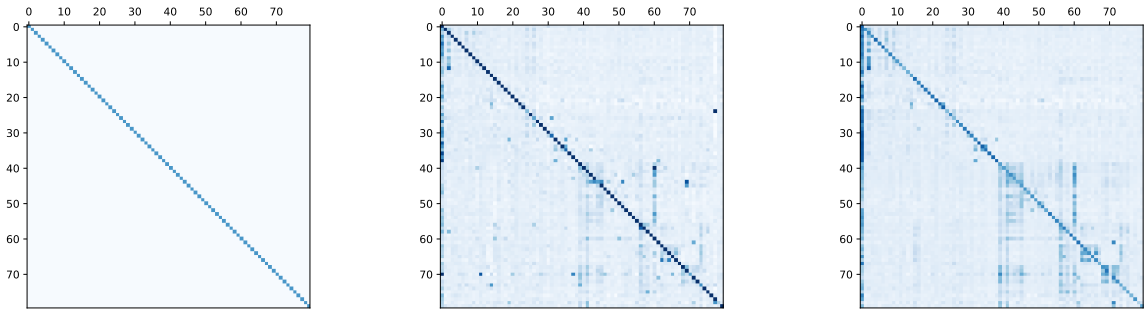
classes. The label vector is affected by noise, hence $\tilde{\boldsymbol{y}}$ can be clean ($\boldsymbol{y}$) or noisy ($\hat{\boldsymbol{y}}$). Similar to GLC [13], we assume that a subset of the data, i.e. *gold dataset* $\mathcal{G} \subset \mathcal{D}$, can be trusted. $|\mathcal{G}|$ contains samples $(\boldsymbol{x}, \boldsymbol{y})$ with no corrupted labels. We refer to the rest of the samples $(\boldsymbol{x}, \hat{\boldsymbol{y}})$ with potentially corrupted labels as *silver dataset* $\mathcal{S} = \mathcal{D} - \mathcal{G}$. We define the *trusted fraction* as the ratio $\frac{|\mathcal{G}|}{|\mathcal{G}|+|\mathcal{S}|}$. Furthermore, we assume that a small dataset of clean single-label images $\mathcal{G}_S$ is available. We use these sets to train a *silver* $f(.;\theta)$ and a *gold* $g(.;\phi)$ classifier and estimate the noise given by a $K \times K$ noise corruption matrix $\mathbf{C}$. The elements $\boldsymbol{C}_{ij}$ are the probability of label $i$ to be flipped into label $j$, formally:

$$\boldsymbol{C}_{ij} = p(\widehat{y}_j = 1 \wedge \widehat{y}_i = 0 | y_j = 0 \wedge y_i = 1)$$

### 3.2 Overview of GALC-SLR

We propose a novel *Gold Asymmetric Loss Correction with Single-Label Regulators* training approach. GALC-SLR combines an asymmetric loss approach with a gold loss correction approach to counter noisy labels. The asymmetric loss treats relevant and irrelevant labels differently and has been shown to obtain impressive results on several MLC datasets [4]. The gold loss correction (GLC) assumes that a small subset of trusted samples is available to accurately estimate the true corruption matrix. It is a powerful method that achieves impressive results under both symmetric and asymmetric noise in the single-label setting [13]. The original GLC assumes conditional independence of $\boldsymbol{y}$ given $\boldsymbol{x}$. This assumption holds when $\boldsymbol{y}$ is deterministic in $\boldsymbol{x}$. This does not hold for multi-label classification, because an image can have multiple labels. Furthermore, in multi-label classification there can be label correlations that the original formula does not take into account, making it impossible to target a specific label. To derive an accurate multi-label, rather than single-label, noise corruption matrix GALC-SLR uses single-label regulators and sigmoid classification which gives a more reliable representation of the noise in the multi-label context.

Figure 3 presents an overview of the GALC-SLR method. It includes three steps. Step 1: we train a classifier $f(.;\theta)$ using ASL loss on the noisy samples in $\mathcal{S}$. Step 2 is the heart of GALC-SLR. We use $f$ with the trusted samples in $\mathcal{G}$ to estimate the noise corruption matrix and correct it via single-label regulators derived via the samples in $\mathcal{G}_S$. Step 3: we



(a) True (b) GALC-SLR (c) GLC

**Figure 4: Comparison between multi-label corruption matrices with 40% noise**

**Algorithm 1** GALC-SLR $\widehat{C}$ estimation

**Input:** Untrusted data $S$, silver classifier $f$
**Input:** Trusted single-label data $G_S$
**Output:** Estimated $\widehat{\boldsymbol{C}}$
*/* Calculate single-label Regulators */*
Fill $Reg \in \mathbb{R}^{K \times K}$ with zeros
*/* For each label, calculate its Regulator row by taking the mean silver **softmax** predictions over the single-label images */*
**for** $k = 1, ..., K$ **do**
  $num\_examples = 0$
  **for** $(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{G}_S$ such that $\boldsymbol{y}_{ik} = 1$ **do**
    $num\_examples \mathrel{+}= 1$
    */* Add silver softmax prediction to kth row */*
    $Reg_{k\bullet} \mathrel{+}= f_{soft}(\boldsymbol{x}_i)$
  **end for**
  $Reg_{k\bullet} \mathrel{/}= num\_examples$
**end for**
*/* Estimate multi-label corruption matrix */*
Fill $\widehat{\boldsymbol{C}} \in \mathbb{R}^{K \times K}$ with zeros
*/* For each label, calculate its corruption row by taking the mean of re-balanced silver **sigmoid** predictions over the untrusted samples */*
**for** $k = 1, ..., K$ **do**
  $num\_examples = 0$
  **for** $(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{S}$ such that $\boldsymbol{y}_{ik} = 1$ **do**
    $num\_examples \mathrel{+}= 1$
    $num\_other\_labels = 0$
    */* Sum up the other labels' regulators */*
    Fill $regulators \in \mathbb{R}^K$ with zeros
    **for** $p = 1, ..., K$ such that $p \neq k$ & $\boldsymbol{y}_{ip} = 1$ **do**
      $num\_other\_labels \mathrel{+}= 1$
      $regulators \mathrel{+}= Reg_{p\bullet}$
    **end for**
    */* Correct sigmoid prediction via regulators */*
    $\widehat{\boldsymbol{C}}_{k\bullet} \mathrel{+}= f_{sig}(\boldsymbol{x}_i) - regulators$
    */* Rebalance towards target label k */*
    $\widehat{\boldsymbol{C}}_{k\bullet} \mathrel{+}= Reg_{k\bullet} * num\_other\_labels$
  **end for**
  $\widehat{\boldsymbol{C}}_{k\bullet} \mathrel{/}= num\_examples$
**end for**
*/* Final scaling */*
$\widehat{\boldsymbol{C}} = sig(\widehat{\boldsymbol{C}})$

train the final classifier $g(.;\phi)$ using ASL loss on samples from $\mathcal{G}$ and corrected samples from $\mathcal{S}$.

### 3.3 Noise Corruption Matrix Estimation

First, we train a *silver classifier* $f(\boldsymbol{x};\theta) = \widehat{p}(\hat{\boldsymbol{y}}|\boldsymbol{x})$ on $\mathcal{S}$, using the asymmetric loss from [4]:

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

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 $\gamma_+, \gamma_-$ are the focusing parameters. Finally, $p_m = \max(p - m, 0)$ denotes the shifted probability by a margin hyperparameter $m$. Given the labels in $\mathcal{S}$ are potentially corrupted, $f$ is not a reliable classifier for our final predictions. However, we can use $f$ to estimate our multi-label corruption matrix $\widehat{\boldsymbol{C}}$.

Algorithm 1 depicts our novel multi-label corruption matrix estimation. First, we calculate the *single-label regulators* by taking the average of our silver softmax predictions for each label (lines 5-15):

$$Reg_{k\bullet} = \frac{\sum_{i=1}^{N} f_{soft}(\boldsymbol{x_i})}{N}, \ \ \forall k \in K$$

where $Reg \subset \mathbb{R}^{K \times K}$ and $Reg_{k\bullet}$ denotes the $k^{th}$ matrix row. This is the main step in our method. It not only allows to target a specific label $k$ for its noise corruption estimation but also to regulate the label correlations from the multi-label images. The next step is to explicitly estimate the noise corruption matrix. For each label $k$, we sum the rows in $Reg$ for the labels which are present in the image except $k$ (lines 26-29). Next, for each row of the noise corruption matrix, we regulate the sigmoid predictions of $f$ by subtracting the summed regulators (line 31). Finally, we rebalance each row using the number of other labels in the image (line 33) and take the average over the number of samples (line 35) scaled via a sigmoid (line 38). Fig. 4 compares the multi-label corruption matrix estimated by GALC-SLR for 40% symmetric label noise against the injected –ground truth– one, and the GLC estimated one. We observe that GALC-SLR's estimation is more resistant to imbalanced data with respect to GLC's estimation. 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. Note that to highlight this effect we avoid the last sigmoid scaling for better contrast in the figures.

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

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

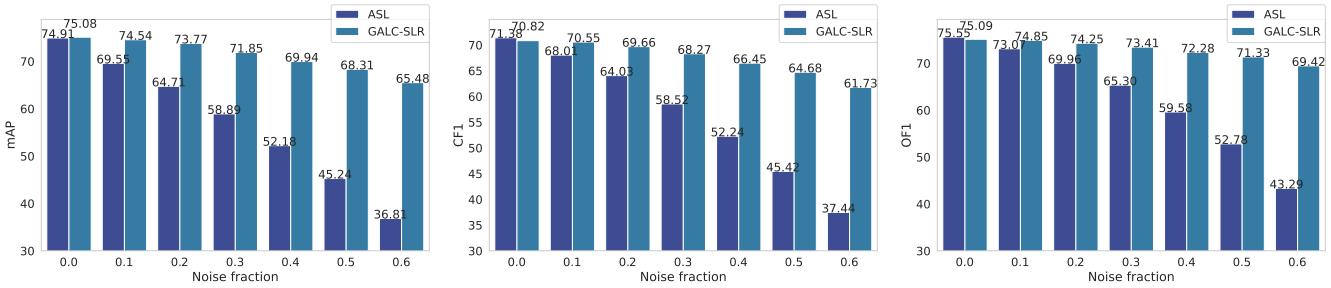
## 4 Evaluation

### 4.1 Experiment setup

**Datasets**. We evaluate GALC-SLR using MS-COCO [21] dataset.

**Table 1: Evaluation results of GALC-SLR and ASL on MS-COCO with symmetric label noise**

| Noise | ASL | | | GALC-SLR | | |
|---|---|---|---|---|---|---|
| | **mAP** | **CF1** | **OF1** | **mAP** | **CF1** | **OF1** |
| *0%* | 74.91 | 71.38 | 75.55 | **75.08** | 70.82 | 75.09 |
| *10%* | 69.55 | 68.01 | 73.07 | **74.54** | 70.55 | 74.85 |
| *20%* | 64.71 | 64.03 | 69.96 | **73.77** | 69.66 | 74.25 |
| *30%* | 58.89 | 58.52 | 65.30 | **71.85** | 68.27 | 73.41 |
| *40%* | 52.18 | 52.24 | 59.58 | **69.94** | 66.45 | 72.28 |
| *50%* | 45.24 | 45.42 | 52.78 | **68.31** | 64.68 | 71.33 |
| *60%* | 36.81 | 37.44 | 43.29 | **65.48** | 61.73 | 69.42 |




(a) Mean-average precision (mAP) (b) Average per-class F1 (CF1) (c) Average overall F1 (OF1)

**Figure 5: Evaluation of GALC-SLR and ASL on MS-COCO with symmetric label noise**

MS-COCO [21] is a popular real-world dataset widely used for multi-label classification evaluation. The training dataset contains 82,081 images, while the validation dataset consists of 40,137 images, at an input resolution of 224. Each image is tagged on average with 2.9 labels belonging to 80 classes. To explicitly test GALC-SLR in the more challenging multi-label setting, we remove all the images with less than 2 labels from both the training and validation dataset, leading finally to 65,268 and 31,739 images respectively. This does not only elicit a more reliable evaluation, but it also allows for the collection of single-label samples to construct $\mathcal{G}_S$. 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. 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 *gold* ($\mathcal{G}$) and *silver* ($\mathcal{S}$) datasets. As base we use 10% as gold data, leading to 6,526 clean samples and 58,742 samples injected with noisy labels.

**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 [15; 23] and inject symmetric noise, but with an extra step. Specifically, we select a fraction $\eta$, i.e. the noise ratio, of labels and flip them to another class with uniform probability. This corresponds to a noise corruption matrix 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. To evaluate how robust GALC-SLR 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 [4; 28; 33]. and have been shown to dramatically decrease with label noise [38]. 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 [25]. 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. In particular, we use the TResNet-M version. Furthermore, the TResNet network has been pre-trained on the ImageNet-21K dataset. This method of transfer learning has been shown to provide better generalizability and significantly increase prediction accuracy [24].

**Baseline**. As baseline, we compare against ASL [4] using the code provided by the authors. GALC-SLR assumes access to a small subset of clean samples.

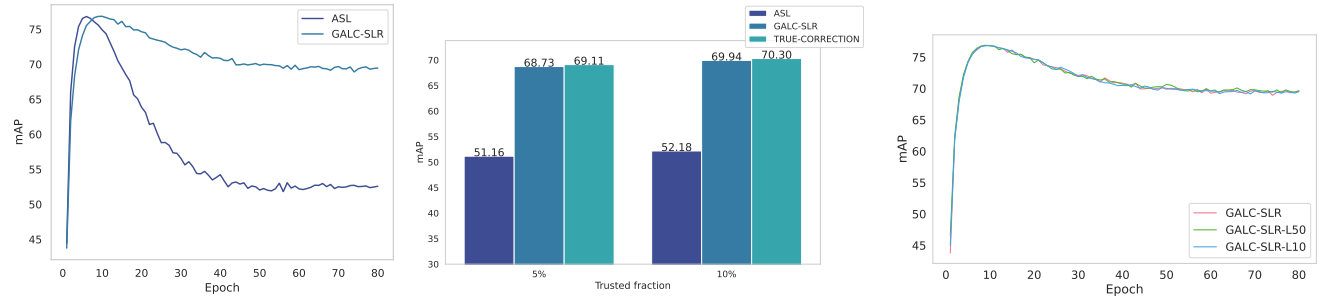
For a fair comparison, we test ASL and GALC-SLR on the same datasets with the same label noise. The only additional knowledge of our method is which labels are trusted, i.e. belonging to the small golden dataset $\mathcal{G}$, and which are potentially corrupted.

**Implementation Details**. We use PyTorch v1.9.0 for both GALC-SLR and ASL, and the default parameters provided in [4] except that we always take 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* [35; 9; 8] benefiting in general from this factor to achieve a better prediction performance in atypical samples. However, [2] 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 GALC-SLR to the performance of ASL under 0% to 60% symmetric noise. We aim to show the effectiveness of our GALC-SLR in robustly learning from noisy data.

Table 1 shows the comparison results. The performance of both systems decreases under increasing noise levels, but GALC-SLR is significantly more robust. In terms of mAP GALC-SLR consistently outperforms ASL for all noise ratios (see Fig. 5(a)). ASL's performance drops an average of 5.34% points with each 10% noise, while GALC-SLR's performance decreases with only 1.07% points. Under severe




(a) Memorization effect under 40% noise (b) Impact of trusted data (c) Impact of single-label data

**Figure 6: Memorization effect and Ablation study of GALC-SLR on MS-COCO with symmetric label noise**

noise, i.e. 60%, the gap between GALC-SLR and ASL is more than 28% points and only 9.6% points worse than without noise. In comparison ASL drops by 38.1% points from 0% to 60% noise. This shows that GALC-SLR is robust even to high noise levels. Similar results apply for both CF1 and OF1, see Fig. 5(b) and Fig. 5(c), respectively. Even if ASL is slightly better in the no-noise case, the performance quickly degrades with additional noise. At 60% GALC-SLR is better by 24.3% and 26.1% points for CF1 and OF1, respectively.

To reliably assess the correctness of GALC-SLR, we also investigate the observed memorization effect for GALC-SLR (depicted in Fig. 6(a)). Both GALC-SLR and ASL follow the same trend. First, they learn the easy patterns, achieving a high accuracy after just a few epochs. However afterward, the performance slowly degrades over training effort and finally stabilizes after 60 epochs. The figure clearly shows the advantage of GALC-SLR over ASL in the different levels at which they plateau. Moreover, one can observe that GALC-SLR has a slight delay in learning at the beginning of the training, i.e. GALC-SLR peaks at epoch 10, while ASL at epoch 6. This observation indicates that GALC-SLR 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. This suggests that our method can also be applied to other existing classifiers and domains.

### 4.3 Ablation Study

To better understand the performance of GALC-SLR, 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 the number of single-label images (studied in experiment II). The base setup of the experiments is the same as in Section 4.1 with the only changes specifically mentioned.

**Experiment I:** Fig. 4 shows visually the difference between the true and our estimated noise corruption matrix. To assess also quantitatively how well our estimation method works, we train classifier $g$ with the true corruption matrix. Fig. 6(b) compares the achieved mAP results under 40% noise. $g$ trained with the true corruption matrix represents the upper performance bound achievable by noise corruption matrix estimators. Fig. 6(b) shows the results across the bars of different colors.

Since GALC-SLR 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 halving the fraction of trusted data, i.e. 5%. This corresponds to 3,263 clean samples and 62,005 samples injected with noisy labels. Fig. 6(b) shows these results via the two different bar plot groups.

With the estimated noise corruption matrix we reach 68.73% and 69.94% mAP and 69.11% and 70.30% using the true noise correction matrix under 5% and 10% of trusted data, respectively. The difference between GALC-SLR and the upper bound (True-Correction) is below 0.5% points in both cases. This shows that our noise estimation is able to capture almost perfectly the impact of the noise corruption matrix and that it works even with a reduced amount of 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. In addition to the previous case using all single-label images, we limit the number of single-label images per class to 50 and 10, referred to as GALC-SLR-L50 and GALC-SLR-L10, respectively. This results in a total of 2,824 for GALC-SLR-L50 and 721 for GALC-SLR-L10 single-label images used.

Fig. 6(c) 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 GALC-SLR. 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, GALC-SLR has a limited dependency on the amount of clean single-label data.

## 5 Responsible Research

In this section, we aim to address any ethical aspects this paper might imply. Due to the lack of human-computer interaction or personal data collection, there are no privacy concerns

to be addressed. Therefore, this section concentrates on the reproducibility concerns of the experiments.

Our method is thoroughly described in Section 3, together with the architecture diagram depicted in Fig. 3. All the used notations have been defined and all the formulas have been clearly explained. Furthermore, the pseudo-code algorithm is provided in Algorithm 1 with clear step-by-step instructions and comments. As an additional reference, Fig. 4 gives a visual representation of the expected noise estimation for our experiment setup.

Section 4 carefully describes all the settings used for the experiments. The dataset is publicly available and all the dataset splits are detailed in Section 4.1. Furthermore, our noise injection method is clearly explained, accompanied by the formula for the corruption matrix calculation. To ensure the reliability of our results, the same random seed has been set along with the experiments, making sure the same dataset splits, as well as the label noise injection, are being used in both our method and the baseline. While the exact numbers might vary in a different experiment, given the added randomness from the symmetric noise injection, all the results should be comparable with the ones reported in this paper.

For further reference, the original code repositories for ASL[2] and GLC[3] are publicly available and our implementation highly relies on the original implementations. We follow conventional settings for the given datasets, all the hyperparameters and training details being either explicitly mentioned in this paper or available in the original works. The additional experiments from the ablation study are also clearly explained in Section 4.3. Moreover, The TResNet-M model that is used throughout all the experiments can be obtained from the ASL ModelZoo[4] archive.

## 6 Conclusion

In this paper, we show the impact of wrong label information on multi-label classification. Motivated by this, we propose the Gold Asymmetric Loss Correction with Single-Label Regulators, a multi-label method that is robust against label noise. This method assumes access to a small set of clean multi-label examples as well as to a small set of clean single-label samples. GALC-SLR uses this additional information in order to accurately model the label noise distribution in a multi-label setting. Through a novel regularization technique that rebalances predictions towards a targeted label, GALC-SLR estimates a noise corruption matrix close to the true matrix. We evaluate GALC-SLR on a real-world dataset under label noise, at multiple corruption levels from low to heavy noise. Results show that GALC-SLR is a powerful method that significantly improves robustness against label noise in multi-label classification.

[2]https://github.com/Alibaba-MIIL/ASL visited June 24, 2021.

[3]https://github.com/mmazeika/glc visited June 24, 2021.

[4]https://github.com/Alibaba-MIIL/ASL/blob/main/MODEL_ZOO.md visited June 24, 2021.

## References


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

[2] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron C. Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *ICML*, volume 70, pages 233–242, 2017.

[3] 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.

[4] Emanuel Ben Baruch, T. Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, M. Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. *ArXiv*, abs/2009.14119, 2020.

[5] 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, 2019.

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

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

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

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

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

[11] 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*, 2018.

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

[13] 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.

[14] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 448–456, 2015.

[15] 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.

[16] 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. *Dataset available from https://github.com/openimages*, 2016.

[17] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). 2009.

[18] Anders Krogh and John Hertz. A simple weight decay can improve generalization. In *NIPS*, volume 4. Morgan-Kaufmann, 1992.

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

[20] 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.

[21] 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.

[22] Eran Malach and S. Shalev-Shwartz. Decoupling "when to update" from "how to update". In *NIPS*, 2017.

[23] 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, 2017.

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

[25] Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben Baruch, Gilad Sharir, and Itamar Friedman. Tresnet: High performance gpu-dedicated architecture. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1400–1409, 2021.

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

[27] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. SELFIE: Refurbishing unclean samples for robust deep learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5907–5915. PMLR, 09–15 Jun 2019.

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

[29] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[30] Lijuan Sun, Songhe Feng, Tao Wang, Congyan Lang, and Yi Jin. Partial multi-label learning by low-rank and sparse decomposition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:5016–5023, 07 2019.

[31] Ryutaro Tanno, A. Saeedi, S. Sankaranarayanan, D. Alexander, and N. Silberman. Learning from noisy labels by regularized estimation of annotator confusion. *CVPR*, pages 11236–11245, 2019.

[32] 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)*, pages 6575–6583, 2017.

[33] Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. Multi-label classification with label graph superimposing. In *AAAI*, pages 12265–12272, 2020.

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

[35] 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.

[36] 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.

[37] 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.

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