The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)

Toward Robustness in Multi-Label Classification: A Data Augmentation Strategy against Imbalance and Noise

Hwanjun Song[1], Minseok Kim[2], Jae-Gil Lee[1]
[1]KAIST
[2]Amazon
{songhwanjun, jaegil}@kaist.ac.kr, kminseok@amazon.com

Abstract

Multi-label classification poses challenges due to imbalanced and noisy labels in training data. We propose a unified data augmentation method, named BalanceMix, to address these challenges. Our approach includes two samplers for imbalanced labels, generating minority-augmented instances with high diversity. It also refines multi-labels at the label-wise granularity, categorizing noisy labels as clean, re-labeled, or ambiguous for robust optimization. Extensive experiments on three benchmark datasets demonstrate that BalanceMix outperforms existing state-of-the-art methods. We release the code at https://github.com/DISL-Lab/BalanceMix.

	Dog	Cat	Bird	Door	Cap	Bowl
Clean Label:	1	0	0	1	0	0
Mislabeling:	0	1	0	1	0	0
Random Flip:	1	1	0	0	0	1
Missing Label:	1					

Figure 1: Examples of noisy labels in multi-label classification. 1 and 0 indicate positive and negative labels, symbolizing the existence of an object class.

Introduction

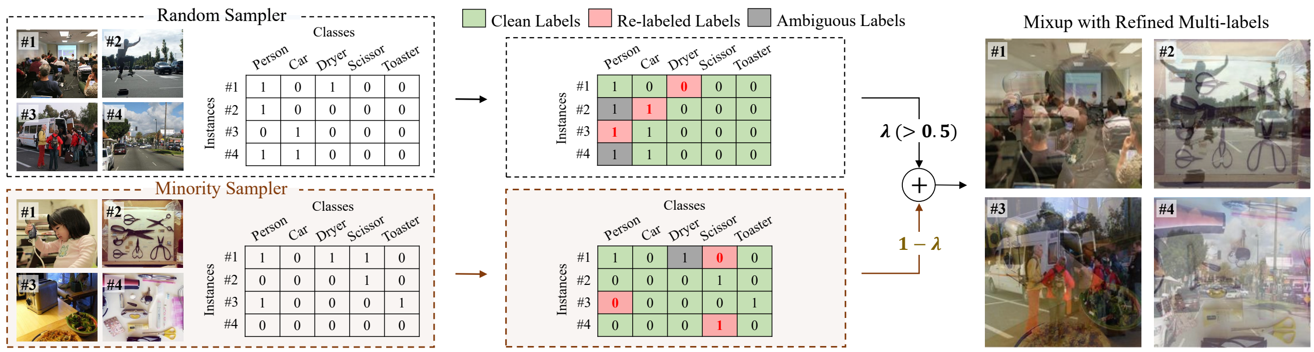
The issue of data-label quality emerges as a major concern in the practical use of deep learning, potentially resulting in catastrophic failures when deploying models in real-world test scenarios (Whang et al. 2021). This concern is magnified in multi-label classification, where instances can be associated with multiple labels simultaneously. In this context, AI system robustness is at risk due to diverse types of data-label issues, although the task can reflect the complex relationships present in real-world data (Bello et al. 2021).

The presence of *class imbalance* occurs when a few majority classes occupy most of the positive labels, and *positive-negative imbalance* arises due to instances typically having fewer positive labels but numerous negative labels. Such imbalanced labels can dominate the optimization process and lead to underemphasizing the gradients from minority classes or positive labels. Additionally, the presence of *noisy labels* stems from the costly and time-consuming nature of meticulous annotation (Song et al. 2022). Labels can be corrupted by adversaries or system failures (Zhang et al. 2020). Notably, instances have both clean and incorrect labels, therefore resulting in diverse cases of noisy labels.

Three distinct types of noisy labels arise in multi-label classification, as illustrated in Fig. 1: *mislabeling*, where a visible object in the image is labeled incorrectly by a human or machine annotator, such as a dog being labeled as a cat; *random flipping*, where labels are randomly flipped by an adversary regardless of the presence of other class objects, such as negative labels for a cat and a bowl being flipped independently to positive labels; and *(partially) missing labels*, where even humans cannot find all applicable class labels for each image, and it is more difficult to detect their absence than to detect their presence (Cole et al. 2021).

Ensuring the robustness of AI systems calls for a *holistic* approach that effectively operates within the following settings: clean, noisy, missing, and imbalanced labels at the same time. However, this task is non-trivial given that minority and noisy labels have similar behavior in learning, e.g., larger gradients, making the task even more complicated and challenging. As a result, prior studies have addressed these two problems *separately* in different setups, assuming either clean or well-balanced training data—i.e., imbalanced clean labels (Lin et al. 2017; Ben-Baruch et al. 2021; Du et al. 2023) and well-balanced noisy labels (Zhao and Gomes 2021; Ferreira, Costeira, and Gomes 2021; Shikun et al. 2022; Wei et al. 2023).

We address this challenge using a new data augmentation method, **BalanceMix**, without complex data preprocessing and architecture change. First, for imbalanced multi-labels, we maintain an additional batch sampler called a *minority sampler*, which samples the instances containing minority labels with high probability, as illustrated in Fig. 2(a). To counter the limited diversity in oversampling, we interpolate the instances sampled from the minority sampler with those sampled from a random sampler using the Mixup (Zhang et al. 2018) augmentation. By mixing with a higher weight to the random instances, the sparse context of the oversampled instances literally *percolates* through the majority of training data without losing diversity. Minority sampling in



(a) Minority Sampling (w. Noisy Multi-labels). (b) Fine-grained Label-wise Management. (c) Mixing Augmentation.

Figure 2: Overview of BalanceMix. Here with MS-COCO, "person" and "car" are the most frequently observed (majority) classes, whereas "hair dryer," "scissor," and "toaster" are the least frequently observed (minority) classes in the training data.

Fig. 2(a) followed by the Mixup augmentation in Fig. 2(c) is called *minority-augmented mixing*.

Then, for noisy multi-labels, we incorporate *fine-grained label-wise management* to feed high-quality multi-labels into the augmentation process. Unlike existing robust learning methods such as Co-teaching (Han et al. 2018) which consider each instance as a candidate for selection or correction, we should move to a finer granularity and consider *each label* as a candidate. As illustrated in Fig. 2(b), the label-wise management step categorizes the entire set of noisy labels into three subsets: *clean* labels which are expected to be correct with high probability; *re-labeled* labels whose flipping is corrected with high confidence; and *ambiguous* labels which need to be downgraded in optimization. Putting our solutions for imbalanced and noisy labels together, BalanceMix is completed, as illustrated in Fig. 2.

Our technical innovations are successfully incorporated into the well-established techniques of oversampling and Mixup, enabling easy integration into the existing training pipeline. Our contributions are threefold: (1) BalanceMix serves as a versatile data augmentation technique, demonstrating reliable performance across clean, noisy, missing, and imbalanced labels. (2) BalanceMix avoids overfitting to minority classes and incorrect labels thanks to minority-augmented mixing with fine-grained label management. (3) BalanceMix outperforms existing prior arts and reaches 91.7mAP on the MS-COCO data, which is the state-of-the-art performance with the ResNet backbone.

Related Work

Multi-label with Imbalance. One of the main trends in this field is solving long-tail class imbalance and positive-negative label imbalance. There have been classical resampling approaches (Wang, Minku, and Yao 2014) for imbalance, but they are mostly designed for a single-label setup. A common solution with multi-labels is the focal loss (Lin et al. 2017), which down-weights the loss value of each label gradually as a model's prediction confidence increases, highlighting difficult-to-learn minority class labels; however, it can lead to overfitting to incorrect labels. The asymmetric focal loss (ASL) (Ben-Baruch et al. 2021) modifies the focal loss to operate differently on positive and negative labels for the imbalance. (Yuan et al. 2023) proposed a balance masking strategy using a graph-based approach.

Multi-label with (Partially) Missing Labels. Annotation in the multi-label setup becomes harder as the number of classes increases. Subsequently, the need to handle missing labels has recently gained a lot of attention. A simple solution is regarding all the missing labels as negative labels (Wang et al. 2014), but it leads to overfitting to incorrect negative ones. There have been several studies with DNNs. (Durand, Mehrasa, and Mori 2019) adopted curriculum learning for pseudo-labeling based on model predictions. (Huynh and Elhamifar 2020) used the inferred dependencies among labels and images to prevent overfitting. Recently, (Cole et al. 2021) and (Kim et al. 2022, 2023) addressed the hardest version, where only a single positive label is provided for each instance. They proposed multiple solutions including label smoothing, and label selection and correction. However, the imbalance problem is overlooked, and all the labels are simply assumed to be clean.

Classification with Noisy Labels. For *single*-label classification, learning with noisy labels has established multiple directions. Most approaches are based on the memorization effect of DNNs, in which simple and generalized patterns are prone to be learned before the overfitting to noisy patterns (Arpit et al. 2017). More specifically, instances with small losses or consistent predictions are treated as clean instances, as in Co-teaching (Han et al. 2018), O2U-Net (Huang et al. 2019), and CNLCU (Xia et al. 2022); instances are re-labeled based on a model's predictions for label correction, as in SELFIE (Song, Kim, and Lee 2019) and SEAL (Chen et al. 2021). A considerable effort has also been made to use semi-supervised learning, as in DivideMix (Li, Socher, and Hoi 2020) and PES (Bai et al. 2021). In addition, a few studies have addressed class imbalance in the noisy single-label setup (Wei et al. 2021; Ding et al. 2022), but they cannot be immediately applied to the multi-label setup owing to their inability to handle the various types of label noise caused by the nature of having both clean and incorrect labels in one instance.

For *multi*-label classification with noisy labels, there has yet to be studied actively owing to the inherent complexity including diverse types of label noise and imbalance. CbMLC (Zhao and Gomes 2021) addresses label noise by proposing a context-based classifier, but its architecture is confined to graph neural networks and requires large pre-trained word embeddings. A method by Hu et al.(Hu et al. 2018) utilizes a teacher-student network with feature transformation. SELF-ML (Ferreira, Costeira, and Gomes 2021) re-labels an incorrect label using a combination of clean labels, but it works only when multi-labels can be defined as attributes associated with each other. ASL (Ben-Baruch et al. 2021) solves the problem of mislabeling by shifting the prediction probability of low-confidence negative labels, making their losses close to zero in optimization. T-estimator (Shikun et al. 2022) solves the estimation problem of the noise transition matrices in the multi-label setting.

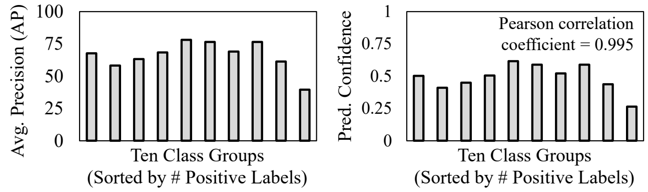
Oversampling with Mixup. Prior studies have applied Mixup to address class imbalance (Guo and Wang 2021; Wu et al. 2020; Galdran, Carneiro, and González Ballester 2021; Park et al. 2022). Yet, they mainly focus on *single*-label classification, overlooking positive-negative imbalances and noisy labels. We propose the first approach that uses predictive confidence to dynamically adjust the degree of oversampling for both types of imbalance while employing label-wise management for noisy labels.

Problem Definition

A multi-label multi-class classification problem requires training data $\mathcal{D}$, a set of two random variables $(\mathbf{x}, \mathbf{y})$ which consists of an instance (d-dimensional feature) $\mathbf{x} \in \mathcal{X} (\subset \mathbb{R}^d)$ and its multi-label $\mathbf{y} \in \{0,1\}^K$, where K is the number of applicable classes. However, in the presence of label noise, the noisy multi-label $\tilde{\mathbf{y}} \in \{0,1\}^K$ possibly contains incorrect labels originated from mislabeling, random flipping, and missing labels; that is, a noisy label $\tilde{y_k} \in \tilde{\mathbf{y}}$ may not be equal to the true label $y_k \in \mathbf{y}$. Thus, let $\tilde{\mathcal{D}} = \{(\mathbf{x}_n, \tilde{\mathbf{y}}_n)\}_{n=1}^N$ be the noisy training data of size N.

Label Noise Modeling. We define three types of label noise. From the statistical view, (1) *mislabeling* is defined as *class-dependent* label noise, where a class object in the image is incorrectly labeled as another class object that may not be visible. The ratio of a class c_1 being mislabeled as c_2 is formulated by $\rho_{c_1 \rightarrow c_2} = p(\tilde{y}_{c_1}=0, \tilde{y}_{c_2}=1 | y_{c_1}=1, y_{c_2}=0)$. In contrast, (2) *random flipping* is *class-independent* label noise, where the presence (or absence) of a class c is randomly flipped with a probability of $\rho_c = p(\tilde{y}_c = 1 | y_c = 0) = p(\tilde{y}_c = 0 | y_c = 1)$, which is independent of the presence of other classes. This scenario can be caused by an adversary's attack or a system failure. Last, (3) *missing labels* from partial labeling can be considered as a type of label noise, where all missing labels are treated as negative ones.

Optimization. To deal with multi-labels in optimization, the most widely-used approach is solving K binary classification problems using the binary cross-entropy (BCE) loss.



(a) Prediction Confidence. (b) Average Precision (AP).

Figure 3: AP and Prediction confidence in COCO using the BCE loss at the 40% of training epochs, where 80 classes are partitioned into ten groups in the descending order of positive label frequency. The Pearson correlation coefficient is computed between ten class groups.

Given a DNN parameterized by Θ, the DNN is updated via stochastic gradient descent to minimize the expected BCE loss on the mini-batch $B \subset \tilde{\mathcal{D}}$,

$$\mathcal{L}(B;\Theta) = \frac{1}{|B|} \sum_{(\mathbf{x},\tilde{\mathbf{y}})\in B} \sum_{k=1}^{K} \mathrm{BCE}(f_{(\mathbf{x},\tilde{y}_k)}), \text{ where} \tag{1}$$
$$\mathrm{BCE}(f_{(\mathbf{x},\tilde{y}_k)}) = -\tilde{y}_k \cdot \log(f_{(\mathbf{x},\tilde{y}_k)}) - (1-\tilde{y}_k) \cdot \log(1-f_{(\mathbf{x},\tilde{y}_k)}).$$

Given the instance $\mathbf{x}$, $f_{(\mathbf{x},\tilde{y}_k)}$ and $1 - f_{(\mathbf{x},\tilde{y}_k)}$ are the confidence in presence and absence, respectively, for the k-th class by the model Θ. BalanceMix is built on top of this standard optimization pipeline for multi-label classification.

Methodology: BalanceMix

Our primary idea is to generate minority-augmented instances and their reliable multi-labels through *data augmentation*. We now detail the two main components, which achieve balanced and robust optimization by *minority-augmented mixing* and *label-wise management*.

Minority-augmented Mixing

To relieve the class imbalance problem, prior studies either oversample the minority class labels or adjust their loss values (Tarekegn, Giacobini, and Michalak 2021). These methods are intuitive but rather intensify the overfitting problem since they rely on a few minority instances with limited diversity (Guo and Wang 2021). On the other hand, we leverage random instances to increase the diversity of minority instances by *separately* maintaining two samplers in Fig. 2.

Confidence-based Minority Sampling. Prior oversampling methods that rely on the frequency of positive labels face two key limitations. First, this frequency alone does not identify the minority multi-labels with low AP values; as illustrated in Fig. 3(a), the class group with few positive labels does not always have lower AP values due to the complexity of the two types of imbalance in the multi-label setup. Second, there is a risk of overfitting because of sticking to the same oversampling policy during the entire training period.

To address these limitations, we first propose to employ the prediction confidence $f_{(\mathbf{x},\tilde{y}_k)}$, which exhibits a strong correlation with the AP, as shown in Fig. 3(b). We opt to oversample the instances with low prediction confidence in their multi-labels, as they are expected to contribute the most significant increase in the AP. Initially, We define two confidence scores for a *specific* class k,

$$\mathrm{P}(k) = \frac{1}{|\mathcal{P}_k|}\sum_{(\mathbf{x},\tilde{\mathbf{y}})\in\mathcal{P}_k} f_{(\mathbf{x},\tilde{y}_k)},\ \ \mathrm{A}(k) = \frac{1}{|\mathcal{A}_k|}\sum_{(\mathbf{x},\tilde{\mathbf{y}})\in\mathcal{A}_k}(1-f_{(\mathbf{x},\tilde{y}_k)}), \tag{2}$$
$$\text{s.t. } \mathcal{P}_k = \{(\mathbf{x},\tilde{\mathbf{y}})\in\tilde{\mathcal{D}}:\tilde{y}_k=1\},\ \mathcal{A}_k = \{(\mathbf{x},\tilde{\mathbf{y}})\in\tilde{\mathcal{D}}:\tilde{y}_k=0\},$$

which are the expected prediction confidences, respectively, for the presence (P) and absence (A) of the k-th class. Next, the confidence score of an instance $(\mathbf{x},\tilde{\mathbf{y}})$ is defined by aggregating Eq. (2) for *all* the classes,

$$\mathrm{Score}(\mathbf{x},\tilde{\mathbf{y}}) = \sum_{k=1}^{K}\mathbf{1}_{[\tilde{y}_k=1]}\mathrm{P}(k) + \mathbf{1}_{[\tilde{y}_k=0]}\mathrm{A}(k). \tag{3}$$

Then, the sampling probability of $(\mathbf{x},\tilde{\mathbf{y}})$ is formulated as

$$p_{\text{sampling}}\big((\mathbf{x},\tilde{\mathbf{y}});\tilde{\mathcal{D}}\big) = \frac{1/\mathrm{Score}(\mathbf{x},\tilde{\mathbf{y}})}{\sum_{(\mathbf{x}',\tilde{\mathbf{y}}')\in\tilde{\mathcal{D}}} 1/\mathrm{Score}(\mathbf{x}',\tilde{\mathbf{y}}')}. \tag{4}$$

By doing so, we consider positive-negative imbalance together with class imbalance by relying on the prediction confidence, which marks a significant difference from existing methods (Galdran, Carneiro, and González Ballester 2021; Park et al. 2022) that considers only the class imbalance of positive labels. Further, our confidence-based sampling dynamically adjusts the degree of oversampling over a training period, mitigating the risk of overfitting. Minority instances are initially oversampled with a high probability, but the degree of oversampling gradually decreases as the imbalance problem gets resolved.

Mixing Augmentation. To mix the instances from the two samplers. We adopt the Mixup (Zhang et al. 2018) augmentation because it can mix two instances even when multi-labels are assigned to them. Let $(\mathbf{x}_R,\tilde{\mathbf{y}}_R)$ and $(\mathbf{x}_M,\tilde{\mathbf{y}}_M)$ be the instances sampled from the random and minority samplers, respectively. The minority-augmented instance is generated by their interpolation,

$$\mathbf{x}^{mix} = \lambda\mathbf{x}_R + (1-\lambda)\mathbf{x}_M,\ \tilde{\mathbf{y}}^{mix} = \lambda\tilde{\mathbf{y}}_R + (1-\lambda)\tilde{\mathbf{y}}_M \\ \text{where } \lambda = \max(\lambda', 1-\lambda'), \tag{5}$$

and $\lambda' \in [0,1] \sim \mathrm{Beta}(\alpha,\alpha)$. Here, random samplers increase the diversity of instances, but lead to imbalanced optimization in the presence of imbalanced labels. In contrast, the random sampler mitigates the imbalance problem by oversampling instances with low confidence, but overfitting occurs due to limited diversity. To maintain high diversity in balanced optimization, by the second row of Eq. (5), λ becomes greater than or equal to 0.5; thus, the instance of the random sampler amplifies diversity, while that of the minority sampler adds the context of minority classes. *Mixing one random instance and one controlled (minor) instance*, instead of mixing two random instances, is a simple yet effective strategy, as shown in the evaluation.

Fine-grained Label-wise Management

Before mixing the two instances by Eq. (5), to make noisy multi-labels reliable in support of robust optimization, we perform *label-wise* refinement.

Clean Labels. To relieve the imbalance problem in label selection, we separately identify clean labels for each class. Let $L_{(\tilde{y}_k=1)}$ and $L_{(\tilde{y}_k=0)}$ be the sets of the BCE losses of the positive and negative labels of the k-th class,

$$L_{(\tilde{y}_k=l)} = \{\mathrm{BCE}(f_{(\mathbf{x},\tilde{y}_k=l)}) \mid (\mathbf{x},\tilde{\mathbf{y}})\in\tilde{\mathcal{D}} \wedge \tilde{y}_k\in\tilde{\mathbf{y}} \wedge \tilde{y}_k = l\}, \tag{6}$$

where l is 1 or 0 for the positive or negative label.

Clean labels exhibit loss values smaller than noise ones due to the memorization effect of DNNs (Li, Socher, and Hoi 2020). Hence, we fit a bi-modal univariate Gaussian mixture model (GMM) to each set of the BCE losses in using the expectation-maximization (EM) algorithm, returning $2\times K$ GMM models for positive and negative labels of K classes,

$$p_{\mathcal{G}} = \{(\mathcal{G}_{(\tilde{y}_k=1)}, \mathcal{G}_{(\tilde{y}_k=0)})\}_{k=1}^{K}. \tag{7}$$

Given the BCE loss of $\mathbf{x}$ for the k-th positive or negative label, its clean-label probability is obtained by the posterior probability of the corresponding GMM,

$$p_{\mathcal{G}}(\mathbf{x},\tilde{y}_k=l) = \frac{\mathcal{G}_{(\tilde{y}_k=l)}(\mathrm{BCE}(f_{(\mathbf{x},\tilde{y}_k=l)})|g)\cdot\mathcal{G}_{(\tilde{y}_k=l)}(g)}{\mathcal{G}_{(\tilde{y}_k=l)}(\mathrm{BCE}(f_{(\mathbf{x},\tilde{y}_k=l)}))}, \tag{8}$$

where g denotes a modality for the small-loss (clean) label. Thus, a label with $p_{\mathcal{G}} > 0.5$ is marked as being clean.

The time complexity of GMM modeling is $\mathcal{O}(NGD) = \mathcal{O}(N)$ and thus linear to the number of instances N, where the number of modalities $G=2$ and the number of dimensions $D=1$ (see (Trivedi et al. 2017) for the proof of the time complexity). Since we model the GMMs once per epoch, the cost involved is expected to be small compared with the training steps of a complex DNN.

Re-labeled Labels. Before the overfitting to noisy labels, a model's prediction delivers useful underlying information on correct labels (Song, Kim, and Lee 2019). Therefore, we modify the given label if it is not selected as a clean one but the model exhibits high confidence in predictions. To obtain a stable confidence from the model, we ensemble the prediction confidences on two augmented views created by RandAug (Cubuk et al. 2020). Given two differently-augmented instances from the original instance $\mathbf{x}$ whose $p_{\mathcal{G}}(\mathbf{x},\tilde{y}_k) \le 0.5$, the k-th label is re-labeled by

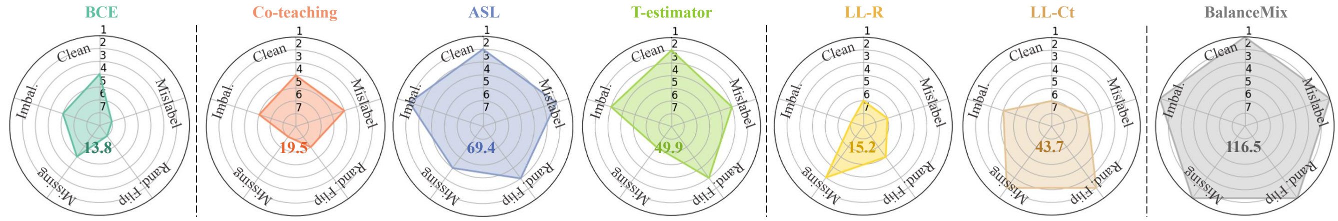
$$1/2\cdot(f_{(\mathrm{aug}_1(\mathbf{x}),\tilde{y}_k)} + f_{(\mathrm{aug}_2(\mathbf{x}),\tilde{y}_k)}) > \epsilon \implies \tilde{y}_k = 1, \\ 1/2\cdot(f_{(\mathrm{aug}_1(\mathbf{x}),\tilde{y}_k)} + f_{(\mathrm{aug}_2(\mathbf{x}),\tilde{y}_k)}) < 1-\epsilon \implies \tilde{y}_k = 0, \tag{9}$$

where ϵ is the confidence threshold for re-labeling.

Ambiguous Labels. The untouched labels, which are neither clean nor re-labeled, are regarded as being ambiguous. These labels are potentially incorrect, but they could hold meaningful information in learning with careful treatment. To squeeze the meaningful information and reduce the potential risk, for these ambiguous labels, we execute *importance reweighting* which decays a loss based on the clean-label probability estimated by Eq. (8).

Optimization with BalanceMix

Given two instances sampled from the random and minority samplers, their multi-labels are first refined by the label-wise management setup. The *reliability* of each refined label is stored as C for clean labels, R for re-labeled labels,



(a) Default. (b) Methods for Noisy Labels. (c) Methods for Missing Labels. (d) Ours.

Figure 4: Performance ranking (1–7) from five different perspectives. BalanceMix is the most versatile to handle diverse types of label issues in multi-label classification. A number in a pentagon is its area, roughly meaning the overall performance.

Class Group		All			Many-shot			Medium-shot			Few-shot		
Category	Method	0%	20%	40%	0%	20%	40%	0%	20%	40%	0%	20%	40%
Default	BCE Loss	83.4	73.1	63.8	86.9	78.4	68.4	84.5	74.0	64.3	64.3	55.5	48.0
Noisy Labels	Co-teaching	82.8	82.5	78.6	87.2	87.0	82.7	84.0	83.7	79.7	61.2	60.9	54.5
	ASL	<u>85.0</u>	<u>82.8</u>	80.3	**88.4**	<u>87.4</u>	<u>85.4</u>	**86.3**	<u>84.0</u>	<u>81.8</u>	<u>67.5</u>	62.5	55.6
	T-estimator	84.3	82.2	<u>80.5</u>	87.5	86.3	85.3	85.4	83.6	81.3	67.0	59.7	<u>61.0</u>
Missing Labels	LL-R	82.5	80.7	75.3	81.0	83.9	79.8	83.8	81.8	76.7	65.7	63.1	52.6
	LL-Ct	79.4	81.3	77.4	72.3	78.8	79.3	80.5	82.6	78.8	67.0	<u>64.3</u>	55.9
Proposed	BalanceMix	**85.2**	**84.3**	**81.6**	**88.4**	**87.7**	**85.5**	<u>86.1</u>	**85.2**	**82.6**	**70.2**	**68.7**	**63.1**

Table 1: Last mAPs on MS-COCO with mislabeling of 0–40%. The 1st and 2nd best values are in bold and underlined.

and U for ambiguous labels. Then, a minority-augmented instance is generated by Mixup with the mixed multi-labels. The reliability of each label of the augmented instance follows that of the instance selected by the random sampler, because it always dominates in mixing by $\lambda \geq 0.5$ in Eq. (5). The loss function of BalanceMix is defined on the minority-augmented mini-batch B_{mix} by

$$\mathcal{L}_{\text{ours}}(B_{mix};\Theta) = \frac{1}{|B_{mix}|}\sum_{(\mathbf{x}^{mix},\tilde{\mathbf{y}}^{mix})\in B_{mix}} \ell(\mathbf{x}^{mix},\tilde{\mathbf{y}}^{mix}),$$
$$\text{where } \ell(\mathbf{x}^{mix},\mathbf{y}^{mix}) = \sum_{k\in\mathsf{C}\cup\mathsf{R}} \text{BCE}(f_{(\mathbf{x}^{mix},\tilde{y}_k^{mix})}) + \sum_{k\in\mathsf{U}} p_{\mathcal{G}}(\mathbf{x}^{mix},\tilde{y}_k^{mix})\cdot\text{BCE}(f_{(\mathbf{x}^{mix},\tilde{y}_k^{mix})}). \quad (10)$$

We perform standard training for warm-up epochs and then apply the proposed loss function in Eq. (10).

Evaluation

Datasets. Pascal-VOC (Everingham et al. 2010) and MS-COCO (Lin et al. 2014) are the most widely-used datasets with well-curated labels of 20 and 80 common classes. In contrast, DeepFashion (Liu et al. 2016) is a real-world in-shopping dataset with noisy weakly-annotated labels for 1,000 descriptive attributes. We use a fine-grained subset of DeepFashion with 16,000 training and 4,000 validation instances as well as multi-labels of 26 attribute classes, which are provided by the authors.

Imbalanced and Noisy Labels. The three datasets contain different levels of natural imbalance. Pascal-VOC, MS-COCO, and DeepFashion have the class imbalance ratios[1] of 14, 339, and 239, and the positive-negative imbalance ratios of 13, 27, and 3, respectively. We artificially contaminate Pascal-VOC and MS-COCO to add three types of label noise. First, for mislabeling, we inject class-dependent label noise. Given a noise rate τ, the presence of the i-th class is mislabeled as that of the j-th class with a probability of $\rho_{i\rightarrow j}$; we follow the protocol used for a long-tail noisy label setup (Wei et al. 2021). For the two different classes i and j,

$$\rho_{i\rightarrow j} = p(\tilde{y}_i=0, \tilde{y}_j=1|y_i=1) = \tau\cdot N_j/(N-N_i), \quad (11)$$

where N_i is the number of positive labels for the i-th class. Second, for random flipping, all positive and negative labels are flipped independently with the probability of τ. Third, for missing labels, we follow the single positive label setup (Kim et al. 2022), where one positive label is selected at random and the other positive labels are dropped.

Algorithms. We use the ResNet-50 backbone pre-trained on ImageNet-1K and fine-tune using SGD with a momentum of 0.9 and resolution of 448×448. We compare BalanceMix with a standard method using the BCE loss (Default) and *five* state-of-the-art methods, categorized into two groups. The former is to handle *noisy* labels based on instance-level selection, loss reweighting, and noise transition matrix estimation—Co-teaching (Han et al. 2018), ASL (Ben-Baruch et al. 2021), and T-estimator (Shikun et al. 2022). The latter is to handle *missing* labels based on label rejection and correction—LL-R and LL-Ct (Kim et al.

[1]The ratio of the number of the instances in the most frequent class to that of the instances in the least frequent class.

Class Group		All			Many-shot			Medium-shot			Few-shot		
Category	Method	0%	20%	40%	0%	20%	40%	0%	20%	40%	0%	20%	40%
Default	BCE Loss	83.4	59.8	43.5	86.9	71.2	65.1	84.5	60.6	43.2	64.3	38.9	30.6
Noisy Labels	Co-teaching	82.8	65.5	43.6	87.2	76.1	61.9	84.0	66.8	43.8	61.2	38.3	25.9
	ASL	<u>85.0</u>	<u>75.0</u>	66.2	**88.4**	<u>84.4</u>	**82.8**	**86.3**	<u>77.0</u>	67.7	<u>67.5</u>	39.2	30.8
	T-estimator	84.3	74.3	<u>69.9</u>	87.5	82.6	80.8	85.4	76.0	<u>71.5</u>	67.4	43.6	39.1
Missing Labels	LL-R	82.5	74.0	69.3	81.0	77.2	76.6	83.8	75.8	71.0	65.7	<u>46.0</u>	38.6
	LL-Ct	79.4	73.2	70.1	72.3	75.5	76.5	80.5	75.0	71.8	67.0	45.5	<u>41.1</u>
Proposed	BalanceMix	**85.2**	**76.5**	**74.5**	**88.4**	**84.5**	<u>81.3</u>	<u>86.1</u>	**78.2**	**76.3**	**70.2**	**46.1**	**43.0**

Table 2: Last mAPs on MS-COCO with random flipping of 0–40%. The 1st and 2nd best values are in bold and underlined.

Datasets		MS-COCO				Pascal-VOC		
Category	Method	All	Many	Medium	Few	All	Medium	Few
Default	BCE Loss	69.7	71.7	70.6	54.4	85.7	89.2	84.2
Noisy Labels	Co-teaching	68.1	61.5	69.2	59.1	80.9	87.2	78.1
	ASL	73.3	**77.7**	74.7	49.2	86.8	82.1	88.8
	T-estimator	16.8	43.2	16.3	3.3	86.2	88.9	85.0
Missing Labels	LL-R	74.2	75.4	75.3	58.7	89.1	<u>91.5</u>	88.1
	LL-Ct	<u>76.9</u>	<u>77.4</u>	<u>78.2</u>	57.6	<u>89.3</u>	<u>91.5</u>	88.3
Proposed	BalanceMix	**77.4**	76.2	**78.5**	**61.3**	**92.6**	**94.5**	**91.8**

Table 3: Last mAPs on MS-COCO and Pascal-VOC in the missing label (single positive label) setup.

2022). For data augmentation, we apply RandAug and Mixup to all methods, except Default using only RandAug. The results of Default with Mixup are presented in Table 5.

As for our hyperparameters, the coefficient α for Mixup is set to be 4.0; and the confidence threshold ϵ for re-labeling is set to be 0.975 for the standard, mislabeling, and random flipping settings with *multiple* positive labels, but it is set to be 0.550 for the missing label setting with a *single* positive label. The source code is available at https://github.com/DISL-Lab/BalanceMix.

Evaluation Metric. We report the overall validation (or test) mAP at the last epoch over three disjoint class subsets: many-shot (more than 10,000 positive labels), medium-shot (from 1,000 to 10,000 positive labels), and few-shot (less than 1,000 positive labels) classes. The result at the last epoch is commonly used in the literature on robustness to label noise (Han et al. 2018).

Overall Analysis on Five Perspectives

Fig. 4 shows the overall performance rankings aggregated[2] on Pascal-VOC and MS-COCO for five different perspectives: "Clean" for when label noise is not injected, "Mislabel" for when labels are mislabeled with the noise ratio of 20–40%, "Rand. Flip" for when labels are randomly flipped with the noise ratio of 20–40%, "Missing" for when the single positive label setup is used, and "Imbal." for when few-shot classes without label noise are used.

[2]For each perspective, we respectively compute the ranking on each dataset and then sum up the rankings to get the final one.

Class Group	All	Many	Medium	Few
BCE Loss	75.2	93.4	84.4	53.4
Co-teaching	66.8	90.7	81.3	32.8
ASL	76.4	94.4	85.2	55.4
T-estimator	75.4	94.7	84.8	53.1
LL-R	75.3	93.3	84.2	53.8
LL-Ct	75.2	92.6	84.2	53.8
BalanceMix	**77.0**	**95.2**	**85.6**	**56.4**

Table 4: mAPs on DeepFashion with real-world noisy multi-labels using seven multi-label classification methods.

Only BalanceMix operates in all scenarios with high performance: its minority-augmented mixing overcomes the problem of *imbalanced labels* while its fine-grained label-wise management adds robustness to *diverse types of label noise*. Except BalanceMix, the five existing methods have pros and cons. The three methods of handling noisy labels in Fig. 4(b) generally perform better for mislabeling and random flipping than the others; but the instance-level selection of Co-teaching is not robust to random flipping where a significant number of negative labels are flipped to positive ones. In contrast, the two methods of handling missing labels in Fig. 4(c) perform better with the existence of missing labels than Co-teaching, ASL, and T-estimator. LL-Ct (label correction) is more suitable than LL-R (label rejection) for mislabeling and random flipping since label correction has a potential to re-label some of incorrect labels. For the imbalance, ASL shows reasonable performance on the few-shot

Component	Clean Label	Mislabel 40%	Rand Flip 40%	Missing Label	Overall (Mean)
Default (BCE Loss)	83.4	63.3	43.0	72.6	65.6
+ Random Sampler ($\approx$ Mixup)	84.2 (+0.8)	67.4 (+3.3)	64.9 (+21.9)	73.3 (+0.7)	72.5 (+6.9)
+ Minority Sampler (in Eq. (4))	85.1 (+1.7)	70.2 (+6.9)	67.2 (+24.2)	74.2 (+1.6)	74.2 (+8.6)
+ Clean Labels (in Eq. (8))	84.9 (−0.2)	76.1 (+5.9)	**74.9** (+7.7)	74.6 (+0.4)	77.6 (+3.4)
+ Re-labeled Labels (in Eq. (9))	**85.3** (+0.4)	80.2 (+3.9)	**74.9** (+0.0)	76.1 (+1.5)	79.1 (+1.5)
+ Ambiguous Labels (in Eq. (10))	**85.3** (+0.0)	**81.6** (+1.4)	74.5 (−0.4)	**77.4** (+1.3)	**79.7** (+0.6)

Table 5: Component analysis of BalanceMix on MS-COCO. The values in parentheses are the gain caused by each component.

subset by adopting a modified focal loss.

Results on Imbalanced and Noisy Labels

We evaluate the performance of BalanceMix on MS-COCO with three types of synthetic label noise and on DeepFashion with *real-world* label noise.

Mislabeling (Table 1). BalanceMix achieves not only the best overall mAP (see the “All” column) with varying mislabeling ratios, but also the best mAP on few-shot classes (see the “Few-shot” column). It shows higher robustness even compared with the three methods designed for noisy labels. ASL performs well among the compared methods, but its weighting scheme of pushing higher weights to difficult-to-learn labels could lead to overfitting to difficult incorrect labels; hence, when the noise ratio increases, its performance rapidly degrades from 67.5% to 55.6% in the few-shot classes. Both methods for missing labels perform better than the default method (BCE), but are still vulnerable to mislabeling.

Random Flipping (Table 2). This is more challenging than mislabeling noise, in considering that even negative labels are flipped by a given noise ratio. Accordingly, the mAP of Co-teaching and ASL drops significantly when the noise ratio reaches 40% (see the “All” column), which implies that instance selection in Co-teaching and loss reweighting in ASL are ineffective to overcome random flipping. T-estimator shows a better result at the noise ratio of 40% than ASL by estimating the noise transition matrix per class. Overall, BalanceMix achieves higher robustness against a high flipping ratio of 40% with fine-grained label-wise management; its performance drops by only $10.7\%p$ (from 85.2% to 74.5%), which is much smaller than $39.2\%p$ (from 82.8% to 43.6%), $18.8\%p$ (from 85.0% to 66.2%), and $14.4\%p$ (from 84.3% to 69.9%) of Co-teaching, ASL, and T-estimator, respectively. Therefore, it maintains the best mAP for all class subsets in general.

Missing Labels (Table 3). Unlike the mislabeling and random flipping, LL-R and LL-Ct generally show higher mAPs than the methods for noisy labels, because LL-R and LL-Ct are designed to reject or re-label unobserved positive labels that are erroneously considered as negative ones. Likewise, the label-wise management of BalanceMix includes the relabeling process, fixing incorrect positive and negative labels to be correct. In addition, it shows higher mAP in the few-shot classes than LL-Ct due to the consideration of imbalanced labels. Thus, it consistently maintains its performance dominance. Meanwhile, T-estimator performs badly in MS-COCO due to the complexity of transition matrix estimation.

Method	Backbone	Resolution	mAP (All)
MS-CMA	ResNet-101	448×448	83.8
ASL	ResNet-101	448×448	85.0
ML-Decoder	ResNet-101	448×448	87.1
BalanceMix	ResNet-101	448×448	**87.4 (+0.3)**
ML-Decoder	TResNet-L	640×640	91.1
ML-Decoder	TResNet-XL	640×640	91.4
BalanceMix	TResNet-L	640×640	**91.7 (+0.6)**

Table 6: State-of-the-art comparison on MS-COCO. The values in parentheses are the improvements over the latest method using the same backbone.

Real-world Noisy Labels (Table 4). A real-world noisy dataset, DeepFashion, likely contains *all* the label noises—mislabeling, random flipping, and missing labels—along with class imbalance. In Table 4, BalanceMix shows the best mAP in all class subsets, while ASL shows the second best mAP. Since Fig. 4 is obtained by aggregating the results from each individual noise type, we confirm that a real-world noisy dataset (e.g., DeepFashion) encompasses these diverse types together. Therefore, our motivation for a holistic approach is of importance for real use cases.

The relatively small performance gain is attributed to a small percentage (around 8%) of noise labels (Song et al. 2022) in DeepFashion, because its fine-grained labels were annotated via a crowd-sourcing platform which can be relatively reliable. The performance gain will increase for datasets with a higher noise ratio.

Component Ablation Study

We conduct a component ablation study by adding the main components one by one on top of the default method. Table 5 summarizes the mAP and average performance of each of five variants. The first variant of using only a random sampler is equivalent to the original Mixup.

First, using only a random sampler like Mixup does not sufficiently improve the model performance, but adding the minority sampler achieves sufficient improvement because it takes imbalanced labels into account. Second, exploiting only the selected clean labels increases the mAP when positive labels are corrupted with mislabeling and random flipping. However, this approach is not that beneficial in the

	Clean Label Selection (C by Eq. (8))				Re-labeling (R by Eq. (9))			
Noise Type	Mislabel 20%		Mislabel 40%		Mislabel 20%		Mislabel 40%	
Training Progress	Precision	Recall	Precision	Recall	Proportion	Accuracy	Proportion	Accuracy
25% Epochs	99.2%	85.3%	96.1%	90.5%	10.1%	98.6%	12.0%	98.9%
50% Epochs	99.0%	88.9%	95.5%	92.7%	9.1%	98.6%	11.2%	98.8%
100% Epochs	98.6%	91.5%	94.5%	94.3%	8.3%	98.5%	9.1%	98.5%

Table 7: Clean label selection and re-labeling performance of BalanceMix on MS-COCO: 2nd-5th columns summarize the label precision and label recall of selecting clean labels, and 6th-9th columns summarize the proportion of re-labeled labels and their re-labeling accuracy.

clean and missing label setups, where all positive labels are regarded as being clean; it also simply discards all (expectedly) unclean negative labels without any treatment. Third, re-labeling complements the limitation of clean label selection, providing additional mAP gains in most scenarios. Fourth, using ambiguous labels adds further mAP improvement except for the random flipping setup.

As all the components generally add a synergistic effect, leveraging all of them is recommended for use in practice.

State-of-the-art Comparison on MS-COCO

We compare BalanceMix with several methods showing the state-of-the-art performance with a ResNet backbone on MS-COCO. The results are borrowed from Ridnik et al. (Ridnik et al. 2023), and we follow exactly the same setting in backbones, image resolution, and data augmentation. BalanceMix is implemented on top of ML-Decoder for comparison. All backbones are pre-trained on ImageNet. The compared methods are developed without consideration of label noise, but we find out that MS-COCO originally has noisy labels.

Table 6 summarizes the best mAP on MS-COCO without synthetic noise injection. For the 448×448 resolution, BalanceMix improves the mAP by 0.3–$0.5\%p$ when using ResNet-101. For the 640×640 resolution, its improvement over ML-Decoder becomes $0.6\%p$ when using TResNet-L. The 91.7mAP of BalanceMix with TResNet-L is even higher than the 91.4mAP of ML-Decoder with TResNet-XL. This record on MS-COCO assures the versatility of BalanceMix for real imbalanced and noisy labels.

Label Precision and Label Recall

The label-wise management of BalanceMix involves selecting clean labels and re-labeling incorrect labels. There are four metrics to evaluate clean label selection and re-labeling performance. Regarding label selection, there are two indicators, label precision and recall, of evaluating how accurate and how many clean labels are chosen from noisy labels, respectively (Han et al. 2018; Song et al. 2022). For convenience, let C be the set of all selected labels from all noisy labels, and L be the set of all clean labels. Then, the label precision and recall are formulated as:

$$\begin{gathered}\text{Label Precision} = |\{\tilde{y} \in \mathsf{C} : \tilde{y} = y\}| \ / \ |\mathsf{C}|, \\ \text{Label Recall} = |\{\tilde{y} \in \mathsf{C} : \tilde{y} = y\}| \ / \ |\mathsf{L}|.\end{gathered} \tag{12}$$

Regarding re-labeling, we evaluate its performance based on the proportion of re-labeled labels and their re-labeling accuracy. Let R be the re-labeled labels among all noisy labels, and D be the entire label in data. Then, the proportion and accuracy are formulated as:

$$\begin{gathered}\text{Relabel Proportion} = |\mathsf{R}| \ / \ |\mathsf{D}|, \\ \text{Relabel Accuracy} = |\{\tilde{y} \in \mathsf{R} : \tilde{y} = y\}| \ / \ |\mathsf{R}|.\end{gathered} \tag{13}$$

Table 7 summarizes their performance on MS-COCO at three different learning progress. For label selection, we evaluate label precision and recall (Han et al. 2018; Song et al. 2022) of the selected clean labels, where they are indicators of how accurate and how many clean labels are chosen, respectively. BalanceMix exhibits very high precision and recall, and the recall increases greatly as training progresses without compromising the precision. For re-labeling, we evaluate the percentage of re-labeled labels and their accuracy. BalanceMix keeps very high re-labeling accuracy in all training phases. Thus, as the model continues to evolve, more clean labels are selected with high precision, and incorrect labels are re-labeled with high accuracy.

Conclusion

We propose BalanceMix, which can handle imbalanced labels and diverse types of label noise. The minority-augmented mixing allows for adding sparse context in minority classes to majority classes without losing diversity. The label-wise management realizes a robust way of exploiting noisy multi-labels without overfitting, categorizing them into clean, re-labeled, and ambiguous labels. Through experiments using real-world and synthetic noisy datasets, we verify that BalanceMix outperforms state-of-the-art methods in each setting of mislabeling, flipping, and missing labels, with the co-existence of severe class imbalance. Overall, this work will inspire subsequent studies to handle imbalanced and noisy labels in a holistic manner.

Acknowledgements

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-00862, DB4DL: High-Usability and Performance In-Memory Distributed DBMS for Deep Learning). Additionally, this work was partly supported by the FOUR Brain Korea 21 Program through the National Research Foundation of Korea (NRF-5199990113928).

References

Arpit, D.; Jastrzebski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A.; Bengio, Y.; et al. 2017. A closer look at memorization in deep networks. In *ICML*, 233–242.

Bai, Y.; Yang, E.; Han, B.; Yang, Y.; Li, J.; Mao, Y.; Niu, G.; and Liu, T. 2021. Understanding and improving early stopping for learning with noisy labels. In *NeurIPS*, 24392–24403.

Bello, M.; Nápoles, G.; Vanhoof, K.; and Bello, R. 2021. Data quality measures based on granular computing for multi-label classification. *Information Sciences*, 560: 51–67.

Ben-Baruch, E.; Ridnik, T.; Zamir, N.; Noy, A.; Friedman, I.; Protter, M.; and Zelnik-Manor, L. 2021. Asymmetric loss for multi-label classification. In *ICCV*, 82–91.

Chen, P.; Ye, J.; Chen, G.; Zhao, J.; and Heng, P.-A. 2021. Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In *AAAI*, 11442–11450.

Cole, E.; Mac Aodha, O.; Lorieul, T.; Perona, P.; Morris, D.; and Jojic, N. 2021. Multi-label learning from single positive labels. In *CVPR*, 933–942.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPRW*, 702–703.

Ding, Y.; Zhou, T.; Zhang, C.; Luo, Y.; Tang, J.; and Gong, C. 2022. Multi-class Label Noise Learning via Loss Decomposition and Centroid Estimation. In *SDM*, 253–261.

Du, Y.; Shen, J.; Zhen, X.; and Snoek, C. G. 2023. SuperDisco: Super-Class Discovery Improves Visual Recognition for the Long-Tail. In *CVPR*, 19944–19954.

Durand, T.; Mehrasa, N.; and Mori, G. 2019. Learning a deep convnet for multi-label classification with partial labels. In *CVPR*, 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*, 88(2): 303–338.

Ferreira, B. Q.; Costeira, J. P.; and Gomes, J. P. 2021. Explainable Noisy Label Flipping for Multi-Label Fashion Image Classification. In *CVPRW*, 3916–3920.

Galdran, A.; Carneiro, G.; and González Ballester, M. A. 2021. Balanced-mixup for highly imbalanced medical image classification. In *MICCAI*, 323–333.

Guo, H.; and Wang, S. 2021. Long-tailed multi-label visual recognition by collaborative training on uniform and re-balanced samplings. In *CVPR*, 15089–15098.

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 *NeurIPS*, 8536–8546.

Hu, M.; Han, H.; Shan, S.; and Chen, X. 2018. Multi-label learning from noisy labels with non-linear feature transformation. In *ACCV*, 404–419.

Huang, J.; Qu, L.; Jia, R.; and Zhao, B. 2019. O2u-net: A simple noisy label detection approach for deep neural networks. In *ICCV*, 3326–3334.

Huynh, D.; and Elhamifar, E. 2020. Interactive multi-label cnn learning with partial labels. In *CVPR*, 9423–9432.

Kim, Y.; Kim, J. M.; Akata, Z.; and Lee, J. 2022. Large Loss Matters in Weakly Supervised Multi-Label Classification. In *CVPR*, 14156–14165.

Kim, Y.; Kim, J. M.; Jeong, J.; Schmid, C.; Akata, Z.; and Lee, J. 2023. Bridging the Gap between Model Explanations in Partially Annotated Multi-label Classification. In *CVPR*, 3408–3417.

Li, J.; Socher, R.; and Hoi, S. C. 2020. DivideMix: Learning with noisy labels as semi-supervised learning. In *ICLR*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *CVPR*, 2980–2988.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common objects in context. In *ECCV*, 740–755.

Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; and Tang, X. 2016. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *CVPR*, 1096–1104.

Park, S.; Hong, Y.; Heo, B.; Yun, S.; and Choi, J. Y. 2022. The Majority Can Help The Minority: Context-rich Minority Oversampling for Long-tailed Classification. In *CVPR*, 6887–6896.

Ridnik, T.; Sharir, G.; Ben-Cohen, A.; Ben-Baruch, E.; and Noy, A. 2023. ML-Decoder: Scalable and versatile classification head. In *WACV*, 32–41.

Shikun, L.; Xiaobo, X.; Hansong, Z.; Yibing, Z.; Shiming, G.; and Tongliang, L. 2022. Estimating Noise Transition Matrix with Label Correlations for Noisy Multi-Label Learning. In *NeurIPS*.

Song, H.; Kim, M.; and Lee, J.-G. 2019. SELFIE: Refurbishing unclean samples for robust deep learning. In *ICML*, 5907–5915.

Song, H.; Kim, M.; Park, D.; Shin, Y.; and Lee, J.-G. 2022. Learning from noisy labels with deep neural networks: A survey. *IEEE TNNLS*.

Tarekegn, A. N.; Giacobini, M.; and Michalak, K. 2021. A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118: 107965.

Trivedi, S. K.; Dey, S.; Kumar, A.; and Panda, T. K. 2017. *Handbook of research on advanced data mining techniques and applications for business intelligence*. IGI Global.

Wang, Q.; Shen, B.; Wang, S.; Li, L.; and Si, L. 2014. Binary codes embedding for fast image tagging with incomplete labels. In *ECCV*, 425–439.

Wang, S.; Minku, L. L.; and Yao, X. 2014. Resampling-based ensemble methods for online class imbalance learning. *IEEE Transactions on Knowledge and Data Engineering*, 27(5): 1356–1368.

Wei, Q.; Feng, L.; Sun, H.; Wang, R.; Guo, C.; and Yin, Y. 2023. Fine-grained classification with noisy labels. In *CVPR*, 11651–11660.

Wei, T.; Shi, J.-X.; Tu, W.-W.; and Li, Y.-F. 2021. Robust long-tailed learning under label noise. *arXiv preprint arXiv:2108.11569*.

Whang, S. E.; Roh, Y.; Song, H.; and Lee, J.-G. 2021. Data Collection and Quality Challenges in Deep Learning: A Data-Centric AI Perspective. *arXiv preprint arXiv:2112.06409*.

Wu, T.; Huang, Q.; Liu, Z.; Wang, Y.; and Lin, D. 2020. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *ECCV*, 162–178.

Xia, X.; Liu, T.; Han, B.; Gong, M.; Yu, J.; Niu, G.; and Sugiyama, M. 2022. Sample selection with uncertainty of losses for learning with noisy labels. In *ICLR*.

Yuan, J.; Zhang, Y.; Shi, Z.; Geng, X.; Fan, J.; and Rui, Y. 2023. Balanced masking strategy for multi-label image classification. *Neurocomputing*, 522: 64–72.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. Mixup: Beyond empirical risk minimization. In *ICLR*.

Zhang, M.; Hu, L.; Shi, C.; and Wang, X. 2020. Adversarial label-flipping attack and defense for graph neural networks. In *ICDM*, 791–800.

Zhao, W.; and Gomes, C. 2021. Evaluating multi-label classifiers with noisy labels. *arXiv preprint arXiv:2102.08427*.