Learning in Imperfect Environment: Multi-Label Classification with Long-Tailed Distribution and Partial Labels

Wenqiao Zhang [1]   Changshuo Liu [2]   Lingze Zeng [2]   Bengchin Ooi [2]   Siliang Tang [1]

Yueting Zhuang [1]

[1] Zhejiang University, China,   [2] National University of Singapore, Singapore,

wenqiaozhang@zju.edu.cn, liu717@comp.nus.edu.sg, Zenglz_pro@163.com, ooibc@comp.nus.edu.sg,

siliang@zju.edu.cn, yzhuang@zju.edu.cn

Abstract

Conventional multi-label classification (MLC) methods assume that all samples are fully labeled and identically distributed. Unfortunately, this assumption is unrealistic in large-scale MLC data that has long-tailed (LT) distribution and partial labels (PL). To address the problem, we introduce a novel task, Partial labeling and Long-Tailed Multi-Label Classification (PLT-MLC), to jointly consider the above two imperfect learning environments. Not surprisingly, we find that most LT-MLC and PL-MLC approaches fail to solve the PLT-MLC, resulting in significant performance degradation on the two proposed PLT-MLC benchmarks. Therefore, we propose an end-to-end learning framework: COrrection → ModificatIon → balanCe, abbreviated as COMIC. Our bootstrapping philosophy is to simultaneously correct the missing labels (Correction) with convinced prediction confidence over a class-aware threshold and to learn from these recall labels during training. We next propose a novel multi-focal modifier loss that simultaneously addresses head-tail imbalance and positive-negative imbalance to adaptively modify the attention to different samples (Modification) under the LT class distribution. In addition, we develop a balanced training strategy by distilling the model's learning effect from head and tail samples, and thus design a balanced classifier (Balance) conditioned on the head and tail learning effect to maintain stable performance for all samples. Our experimental study shows that the proposed COMIC significantly outperforms general MLC, LT-MLC and PL-MLC methods in terms of effectiveness and robustness on our newly created PLT-MLC datasets. Codes and benchmarks are available on the link https://https://github.com/wannature/COMIC

1. Introduction

In recent years, the development of deep learning prosperity to the field of computer vision [20, 17, 19, 16, 23, 22, 25, 24, 43, 42, 18]. Images typically contain multiple objects and concepts, highlighting the importance of multi-label classification (MLC) [34] for real-world tasks. Along with the wide adoption of deep learning, recent MLC approaches have made remarkable progress in visual recognition [36, 38], but the performance is limited by two common assumptions: *all categories have comparable numbers of instances* and *each training instance has been fully annotated with all the relevant labels*. While this conventional setting provides a perfect training environment for various studies, it conceals a number of complexities that typically arise in real-world applications: **i) Long-Tailed (LT) Class Distribution.** With the growth of digital data, the crux of making a large-scale dataset is no longer about where to collect, but how to balance it [33]. However, the cost of expanding the dataset to a larger class vocabulary with balanced data is not linear — but exponential — since the data is inevitably long-tailed following Zipf's distribution [30]. **ii) Partial Labels (PL) of Instances.** In the case of a large number of categories, it is difficult and even impractical to fully annotate all relevant labels for each image [39, 46]. Intuitively, humans tend to focus on different aspects of image contents due to *human bias*, *i.e.,*, their preference, personality and sentiment [41], which indirectly affects how and what we annotate. In fact, LT and PL are often co-occurring, and therefore, the MLC model must be sufficiently robust to handle different data distributions and imperfect datasets.

In this paper, we present a new challenge for MLC at scale, Partial labeling and Long-Tailed Multi-Label Classification (PLT-MLC), with concomitant existence of both PL setting [39] and LT distribution [33] problems. As captured in the overview of PLT-MLC in Figure 1 (a), it has the following three challenges: **i) False Negative Training.** Under the PL setting, the MLC model treats the un-annotated labels (⬤) as *negatives* (⬤), which may produce sub-optimal decision boundary as it adds noise of false *negative* labels (Figure 1 (b)). The situation is further exacerbated in the
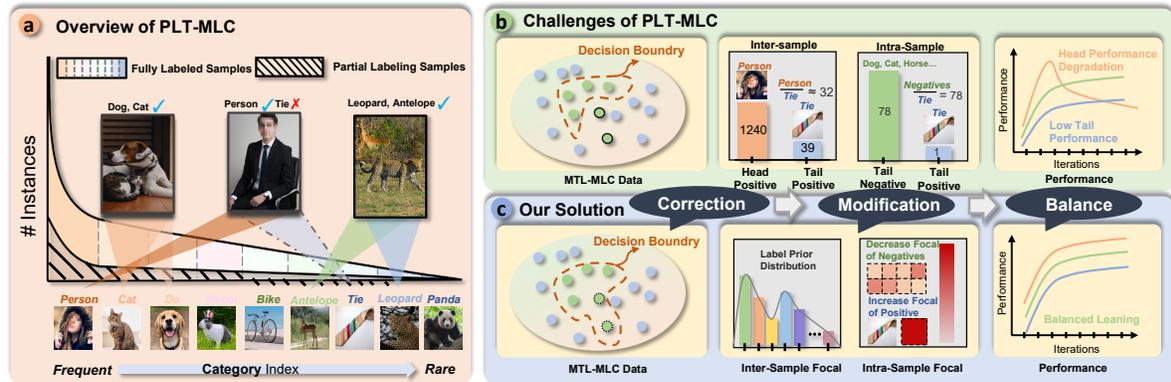
Figure 1: (a) illustrates an overview of the proposed PLT-MLC task. (b) depicts three key challenges of a PLT-MLC task. (c) depicts a concise version of our proposed model for facilitating the PLT-MLC. (● : *positive*, ● : *negative*, ○ : false *negative*, ◉ : corrected *positive*)

LT class distribution as some *tail* categories are prone to missing annotations in practice. For instance, in Figure 1 (a), "person" is the *head* class in the PLT-MLC dataset and is often notable in an image to labeling for annotators. In contrast, the "tie" often occupies a tiny region in the scene compared with the "person". The annotator may miss the "tie" object, which will aggravate the LT distribution and further increase the difficulty of learning from *tail* classes. **ii) Head-Tail and Positive-Negative Imbalance.** There are two imbalance issues in a PLT-MLC task: inter-instance *head-tail* imbalance and intra-instance *positive-negative* imbalance. As shown in Figure 1 (b), the inter-instance ratio of *head positive* (●) "person" (👩) : *tail positive* (●) "tie" (👔) ≈ 32 under the LT data distribution, and the intra-instance ratio of *tail negative* (●) categories : *tail positive* (●)"tie" (👔) = 78 as an image only contains a small fraction of the *positive* labels. Consequently, a robust PLT-MLC model should address the co-occurring imbalances simultaneously. **iii) Head Overfitting and Tail Underfitting.** Different from the general LT distribution, the classification model downplays the minor *tail* and overplays the major *head*. The PLT-MLC has an extreme LT distribution and Figure 1(c) illustrates an interesting phenomenon of MLC model learning: the general classification model is prone to overfitting to *head* class with extensive samples and underfitting to *tail* classes with a few samples. This figure also indicates that only the *medium* class shows a steady growth in performance, which means that existing LT methods focusing on lifting up the *tail* performance may not solve the PLT-MLC problem satisfactorily.

Suppose a trained model is used to correct the missing labels and then an LT classifier is trained using the updated labels, we might not be able to obtain a satisfying PLT-MLC performance, either. While machine learning methods can easily detect the *head* samples, they may have difficulty in identifying the *tail* samples. As a result, the corrected labels

may still exhibit an LT distribution that inevitably hurts balanced learning. Moreover, even when a general LT classifier affords the trade-off to improve the *tail* performance conditioned on the *head* performance drop, it is still incapable of simultaneously addressing the issue of *head* overfitting and *tail* underfitting problem. Further, the decoupled learning paradigm is impractical since it needs the "stop" training and human "re-start" training, *i.e.*, an end-to-end learning scheme is more desirable. Thus, these limitations motivate us to reconsider the solution for the PLT-MLC task.

To this end, we propose an end-to-end PLT-MLC framework: **CO**rrection → **M**odificat**I**on → balan**C**e (Figure 1), called **COMIC**, which progressively addresses the three key PLT-MLC challenges. **Step 1:** The *Correction* module aims to gradually correct the missing labels according to the predicted confidence and dynamically adjusts the classified loss of the corrected samples under the real-time estimated class distribution. **Step 2:** After the label correction, the *Modification* module introduces a novel Multi-Focal Modifier (MFM) Loss, which contains two focal factors to address the two imbalance issues in PLT-MLC independently. Motivated by [3], the first is an intra-instance *positive-negative* factor that determines the concentration of learning on hard *negatives* and *positives* with different exponential decay factors. The second is an inter-instance *head-tail* factor that increases the impact of rare categories, ensuring that the loss contribution of rare samples will not be overwhelmed by frequent ones. **Step 3:** Finally, the *Balance* module measures the model's optimization direction with a calculated moving average vector of the gradient over all past samples. And thus, we devise a *head* model and a *tail* model by subtracting or adding this moving vector, which can respectively improve *head* and *tail* performance. Subsequently, a balanced classifier deduces a balanced learning effect under the supervision of the *head* classifier and *tail* classifier. It protects the model training from being too *medium biased*, and hence

the balanced classifier is able to achieve the balanced learning schema. Notably, our solution is an end-to-end learning framework, which is re-training-free and effectively enables balanced prediction.

Our contributions are three-fold: (1) We present a new challenging task: Partial labeling and Long-Tailed Multi-Label Classification (PLT-MLC), together with two newly designed benchmarks: PLT-COCO and PLT-VOC. (2) We propose an end-to-end PLT-MLC learning framework, called COMIC, to effectively perform the PLT-MLC task as a progressive learning paradigm, *i.e.*, *Correction → Modification → Balance*. (3) Through an extensive experimental study, we show that our method improves all the prevalent LT and ML line-ups on PLT-MLC benchmarks by a large margin.

2. Related Works

Long-Tailed MLC. [37] is the first work that addresses the LT-MLC by extending the re-balanced sampling and cost-sensitive re-weighting methods. It proposes an optimized DB Focal method, which does improve the recognition performance of *tail* classes. Later work, [10] performs uniform and re-balanced samplings from the same training set. Then a two-branch network is developed to enforce the consistency between two branches for collaborative learning on both uniform and re-balanced samplings. However, the above works require careful data initialization, *i.e.*, re-sampling, which is undesirable in practice. Moreover, these methods have not yet considered the missing labeling case, which may not sufficiently deal with the PLT-MLC.

MLC with Partial Labels. Multi-label tasks often involve incomplete training data, hence several methods have been proposed to solve the problem of multi-label learning with missing labels. A simple solution is to treat the missing labels as *negative* labels [31, 21, 4]. However, performance will drop because a lot of ground-truth *positive* labels are initialized as *negative* labels [12]. Current works on PL-MLC mainly focus on the design of networks and training schemes. The common practice is to utilize the customized networks to learn label correlations or classification confidence to realize correct recognition of missing labels [7, 44]. However, the corrected labels learned from a trained model are imbalanced due to the previous training dataset having an LT distribution. Using such recall labels will aggravate the LT distribution in the PLT-MLC dataset and result in an imbalanced performance.

Other Works. [40] learns to generate concept-invariant samples in the environment with semantic gap, which enables the model to classify the samples through causal intervention, yielding improved generalization guarantees; [18] collaborates a cascade of foundation models, thus learning superior interleaved multi-modal instruction-following ability from imperfect supervision.

3. Methodology

This section describes the proposed PLT-MLC framework. We will present each module and its training strategy.

3.1. Problem Formulation

Before presenting our method, we first introduce some basic notions and terminologies. We consider a partially annotated MLC dataset contains C classes and N i.i.d training samples $\mathcal{S} = \{(\mathcal{I}^{(1)}, y^{(1)}), \cdots, (\mathcal{I}^{(N)}, y^{(N)})\}$, where $\mathcal{I}^{(i)}$ denote i^{th} image and label $y^{(i)} = [y_1^{(i)}, \cdots, y_c^{(i)}] \in \{0, 1\}^C$. For a given i^{th} example and category c, $y_c^{(i)} = 0, 1$ respectively means the category is unknown and present. Our proposed COMIC solves the PLT-MLC problem in an end-to-end learning manner: *Correction → Modification → Balance*, with *Reflective Label Corrector* (RLC, in Sec.3.2), *Multi-Focal Modifier* (MFM, in Sec.3.3) and *Head-Tail Balancer* (HTB, in Sec.3.3), as illustrated in Figure 2.

These three modules are designed to seek a balanced model $\mathcal{M}_b(\cdot; \Theta_b)$, parameterized by Θ_b, to predict the presence or absence of each class given an input image. We denote $p = [p_1, \cdots, p_c]$ as the class prediction, computed by the model: $p_c = \sigma(z_c)$ where σ is the sigmoid function, and z_c is the output logit corresponding to class c. The optimized goal of COMIC can be defined as follows:

$$\underbrace{\mathcal{L}((\mathcal{S}); \Theta_b)}_{\text{COMIC Loss}} = \underbrace{\lambda_c \cdot \mathcal{L}_{rlc}}_{\text{RLC Loss}} + \underbrace{\lambda_m \cdot \mathcal{L}_{mfm}}_{\text{MFM Loss}} + \underbrace{\lambda_b \cdot \mathcal{L}_{htb}}_{\text{HTB Loss}} \quad (1)$$

where $\mathcal{L}_{rlc}$, $\mathcal{L}_{mfm}$ and $\mathcal{L}_{htb}$ denote the loss of RLC, MFM and HTB, respectively. λ_c, λ_m and λ_b are hyperparameters.

3.2. Reflective Label Corrector

Reflective Label Corrector (RLC) presents a real-time label correction method for missing labels to alleviate the effect of partially labeled samples. The core idea is to examine the *label likelihood* p of each training image and recall the labels with convinced prediction confidence during training. Interestingly, we found that the model can distinguish a large number of missing labels with high prediction confidence in the early training stage, which implies that we can recall these missing labels during training to boost PTL-MLC learning. Here, we first define a threshold τ and then check the input sample's label likelihood p to check whether it is greater than τ and then calculate the average category possibility P_c of past trained data with class c. If predicted probabilities p_c are highly confident, *i.e.*, $p_c > \max\{\tau, P_c\}$, we regard that the sample misses the label of class c and set a pseudo-label $\hat{y}_c$.

$$\hat{y}_c = \begin{cases} 1, & \text{if } p_c > \max\{\tau, P_c\}, y_c = 0 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Thus, the loss of RLC module, *i.e.*, $\mathcal{L}_{rlc}$ utilizes the MFM loss (refer the details in Sec. 3.3) with these corrected label
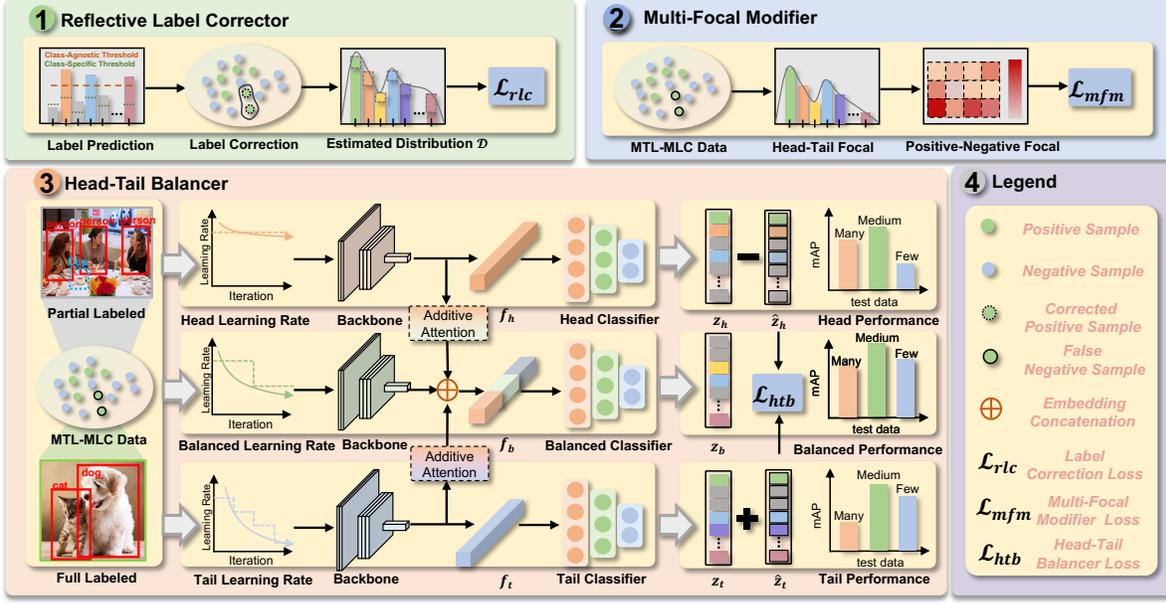
Figure 2: **Overview of COMIC**. RLC module (*Correction*) corrects the missing labels along with the training and dynamically re-weights the sample weight according to the estimated class distribution. MFM module (*Modification*) adjusts the focal of different instances according to *head-tail* and *positive-negative* imbalance under the extreme LT distribution. HTB module (*Balance*) measures the model's optimization direction and correspondingly develops a balanced learning scheme to produce stable PLT-MLC performance.

$\hat{y}$ for model training:

$$\mathcal{L}_{rlc}(p) = \begin{cases} \mathcal{L}_{mfm}^+(p), & \text{if } \hat{y} = 1 \\ \mathbb{1}_{(y=1)} \mathcal{L}_{mfm}^+(p) + \mathbb{1}_{(\hat{y}=0)} \mathcal{L}_{mfm}^-(p), & \text{otherwise} \end{cases}$$
(3)

where $\mathcal{L}_{mfm}^+(p)$ and $\mathcal{L}_{mfm}^-(p)$ refer to the generalized MFM loss function for *positives* and *negatives*. Notably, we find that most of the corrected labels belong to *head* class, which may aggravate the LT distribution. To address this issue, we dynamically adjust the inter-sample *head-tail* factor $\gamma_{ht}^{(i)}$ (details of this factor are explained in Sec. 3.3) according to the dynamic class distribution $\mathcal{D}_t$, which can increase the focal weight for *tail* samples. In addition, we also multiply $\mathcal{L}_{rlc}(p)$ by a coefficient $\frac{\mathcal{B}_s}{\mathcal{N}_t}$ in each training batch to constrain the loss value, where $\mathcal{B}_s$ is the batch size and $\mathcal{N}_t$ is the number of corrected labels.

Through this, RLC can gradually and dynamically correct the potential missing labels during training, which efficiently improves the classifier's performance with recalled labels.

3.3. Multi-Focal Modifier

We first revisit the focal loss [26], which is a widely-used solution in the *positive-negative* imbalance problem. It redistributes the loss contribution of easy samples and hard samples, which greatly weakens the influence of the majority

of *negative* samples.

$$\mathcal{L}_{fl}(p) = \begin{cases} \mathcal{L}_{fl}^+ = (1-p)^\gamma \log(p), & \text{if } y = 1 \\ \mathcal{L}_{fl}^- = p^\gamma \log(1-p), & \text{if } y = 0 \end{cases}$$
(4)

where γ is the focusing parameter, $\gamma = 0$ yields binary cross-entropy. By setting $\gamma > 0$, the contribution of easy *negatives* (with low probability, $p \ll 0.5$) can be down-weighted in the loss, enabling the model to focus on harder samples.

However, the focal loss may not satisfactorily resolve the PLT-MLC problem due to two key aspects:

- **Tail Positive Gradient Elimination.** When using focal loss for multi-label training, there is an inner trade-off: high γ sufficiently down-weights the contribution from easy *negatives*, but may eliminate the gradients from the *tail positive* samples [3].

- **Head-Tail Imbalance.** Imbalance among the *positive* categories also exists in MLC, *i.e.*, *head positive-tail positive* imbalance. Rare categories suffer more from severe imbalance issues than frequent ones.

Thus, we propose a Multi-Focal Modifier (MFM) loss that decouples γ at two granularities of focal factors, *i.e.*, an intra-sample *positive-negative* (P-N) factor γ_{pn} and an

inter-sample *head-tail* (H-T) factor γ_{ht}.

$$\gamma^{(i)} = \begin{cases} \gamma^{(i)+} = \gamma_{pn}^+ + w^+ \cdot \gamma_{ht}^{(i)}, & \text{if } y = 1 \\ \gamma^{(i)-} = \gamma_{pn}^- + w^- \cdot \gamma_{ht}^{(i)}, & \text{if } y = 0 \end{cases} \quad (5)$$

where $\gamma^{(i)+}$ and $\gamma^{(i)-}$ control the focal of samples with i^{th} class. Similar to [3], γ_{pn}^+ and γ_{pn}^- decouple the original decay rate γ as two factors, which respectively control the focal of the *positive* and *negative* samples. Since we are interested in emphasizing the contribution of *positive* samples, we set $\gamma_{pn}^- \geq \gamma_{pn}^+$. We achieve better control over the contribution of *positive* and *negative* samples through the designed loss function, which assists the network to learn meaningful features from *positive* samples, despite their rarity. Another focal factor $\gamma_{ht}^{(i)}$ is a variable parameter (≥ 1) associated with the imbalance degree of the ith class. A bigger value of $\gamma_{ht}^{(i)}$ will increase the weight of *tail* samples to encourage the model to pay more attention to the *positive tail* samples, and vice versa. w^+ and w^- are the coefficients that adjust the weight at a fine-grained level. The $\gamma_{ht}^{(i)}$ is the static class distribution $\mathcal{D}$ of training set with max normalization function $\psi(\cdot)$ [29] to adjust the *head-tail* focal.

After applying the decoupled $\gamma^{(i)+}$ and $\gamma^{(i)-}$ into our MFM loss, we obtain the loss function as follows (more discussions in the Appendix.):

$$\mathcal{L}_{mfm}(p) = \begin{cases} \mathcal{L}_{mfm}^+ = \sum_{i=1}^C (1-p)^{\gamma^{(i)+}} \log(p), & \text{if } y = 1 \\ \mathcal{L}_{mfm}^- = \sum_{i=1}^C p^{\gamma^{(i)-}} \log(1-p), & \text{if } y = 0 \end{cases}$$
$$(6)$$

By doing so, the MFM module utilizes the multi-grained focal to alleviate the two imbalance problems in the PLT-MLC task, yielding better classification results.

3.4. Head-Tail Balancer

As discussed in the introduction, the extreme LT dataset with numerous *head* samples and a small number of *tail* samples result in a *head* overfitting and *tail* underfitting learning effect. Only the *medium* samples present a superior performance during training, which fails to obtain the balanced performance for the overall samples. To address this issue, we develop a balanced strategy that measures the balanced learning effect under the supervision of the *head* classifier and *tail* classifier to achieve balanced results. Before we delve into the balanced learning, we first measure the moving average of the gradient in the SGD-based optimizer [32]:

$$\mathbf{e}_t = \mu \cdot \mathbf{e}_{t-1} + \text{sum}(g_t), \forall t = 1, \cdots, T. \quad (7)$$

where $\text{sum}(g_t)$ is the accumulated gradient at iteration t, μ is the momentum decay. The average moving vector $\mathbf{e}_t$ records the model's optimization tendency by $\mathbf{e}_{t-1}$ and $\text{sum}(g_t)$.

In our empirical study, we observe that only the *medium* samples obtain a stable learning effect from the early to late

training stage, mainly due to the extreme LT distribution. To simulate the learning effect towards *head/tail* samples, as depicted in Figure 2(c), we reduce/add the moving vector $\mathbf{e}_t$ at each step in *head* model $\mathcal{M}_h$ and *tail* model $\mathcal{M}_t$ respectively, to assist the balanced model $\mathcal{M}_b$ for balanced learning. Notably, we set different learning rate decays for each model to further explore the balanced learning effect. The three models are parallel-trained with their own backbone and classifier. In the feature learning stage, we develop an additive attention [1] that computes the relevance of balanced features $\hat{\mathbf{f}}_b$, and *head* features $\mathbf{f}_h$, *tail* features $\mathbf{f}_t$ extracted from corresponding backbones.

$$\mathbf{f}_b = \text{Attn}(\hat{\mathbf{f}}_b, [\mathbf{f}_h, \mathbf{f}_t])) + \hat{\mathbf{f}}_b \quad (8)$$

where $Attn(\cdot)$ is the additive attention mechanism.

Then updated $\mathbf{f}_b$, $\mathbf{f}_h$, $\mathbf{f}_t$ are input to their classifiers to obtain their logits. We develop the multi-head classifier with normalization [9, 28, 33], which has already been embraced by various methods of empirical practice. The multi-head strategy [35] equally divides the channel of weights and features into N_g groups, which can be considered as N_g times of fine-grained sampling.

$$\mathbf{z}_x = \frac{\rho}{N_g} \sum_{k=1}^{N_g} \frac{w_k^\top \mathbf{f}_x}{(||w_k|| + \eta)||\mathbf{f}_x||}, x \in \{h, t, b\} \quad (9)$$

where ρ is a scaling factor akin to the inverse temperature in Gibbs distribution, η is a class-agnostic baseline energy. w_k is the kth learned parameter matrix.

Subsequently, we measure the *head* and *tail* learning effect by subtracting and adding the average moving vector $\mathbf{e}_t$ to the logits of *head* model and *tail* model, respectively:

$$\hat{\mathbf{z}}_x = \mathbf{z}_x \pm \frac{\rho}{N_g} \sum_{k=1}^{N_g} \frac{\text{sim}(\mathbf{z}_x, \mathbf{e}_t) \cdot (w_j)^\top \mathbf{e}_t}{||w_k|| + \eta}, x \in \{h, t\} \quad (10)$$

where $sim(\cdot, \cdot)$ measures the cosine similarity of vectors.

After obtaining the logits of $\hat{\mathbf{z}}_h$, $\hat{\mathbf{z}}_t$ and $\mathbf{z}_b$, the balanced learning effect needs to distill the *head* and *tail* knowledge from $\hat{\mathbf{z}}_h$ and $\hat{\mathbf{z}}_t$ to enable the stable learning for all samples. Hence, we develop the *head-tail* loss:

$$\mathcal{L}_{htb} = \kappa_h \cdot \mathcal{L}(\phi(\hat{\mathbf{z}}_h) \cdot \phi(\mathbf{z}_b)) + \kappa_t \cdot \mathcal{L}(\phi(\hat{\mathbf{z}}_t) \cdot \phi(\mathbf{z}_b)) \quad (11)$$

where $\mathcal{L}$ is $\mathcal{L}_{mfm}$ and $\phi(\cdot)$ is softmax function. κ_h and κ_t are adaptive weights for *head* and *tail* learning that calculated by $\kappa_h = \frac{(\mathcal{L}(\hat{z}_h))^\alpha}{(\mathcal{L}(\hat{z}_t))^\alpha + (\mathcal{L}(\hat{z}_h))^\alpha}$ and $\kappa_t = \frac{(\mathcal{L}(\hat{z}_t))^\alpha}{(\mathcal{L}(\hat{z}_t))^\alpha + (\mathcal{L}(\hat{z}_h))^\alpha}$, respectively. α is a scaling factor and we study its effect in Sec. 4.3. Such loss can be regarded as the empirical risk minimization (ERM) [6], which adaptively distills the knowledge from the *head* and *tail* models, enabling the balanced model is not biased to *medium* samples and produces a balanced learning effect for the PLT-MLC task.

Table 1: **Performance comparison of the proposed method and baselines on PLT-MLC datasets** (`PLT-COCO` and `PLT-VOC`). $E2E^*$ indicates that the PLT model is learned in an end-to-end manner. A larger score has better performance. Improv. indicates performance improvement. Acronym notations of baselines can be found in Sec. 4.1. We color each row as the ==best==, ==second best== and ==lowest score==.

Category	Methods	E2E*	PLT-COCO Dataset				PLT-VOC Dataset			
			Many Shot	*Medium Shot*	*Few Shot*	*Total Shot*	*Many Shot*	*Medium Shot*	*Few Shot*	*Total Shot*
MLC	BCE [45]	✓	42.57±0.11	56.67±0.19	46.40±0.60	48.92±0.23	67.37±0.18	88.27±0.39	83.79±0.41	78.79±0.14
	Focal [26]	✓	41.05±0.07	58.33±0.12	53.58±0.31	51.39±0.15	67.02±0.11	87.49±0.18	82.82±0.78	78.13±0.23
	ASL [3]	✓	41.60±0.17	58.15±0.15	52.67±0.17	51.20±0.08	67.67±0.10	87.79±0.13	82.23±0.55	78.35±0.11
LT-MLC	DB [37]	✓	44.83±0.31	58.96±0.24	53.82±0.47	52.16±0.36	69.22±0.28	88.56±0.42	83.72±0.35	78.86±0.23
	DB-Focal [37]	✓	45.76±0.25	59.74±0.21	53.85±0.16	52.57±0.27	68.96±0.22	88.89±0.18	83.42±0.20	78.90±0.26
	LWS [13]	-	44.86±0.58	58.79±0.63	53.48±0.51	52.86±0.60	69.08±0.44	88.24±0.55	83.46±0.47	78.28±0.49
PL-MLC	Pseudo-Label [15]	-	41.41±0.41	57.46±0.35	53.12±0.33	51.67±0.37	67.38±0.24	87.58±0.35	83.26±0.42	78.32±0.30
	ML-GCN [5]	✓	43.43±0.53	58.46±0.61	53.74±0.48	52.14±0.55	68.46±0.44	88.17±0.61	82.46±0.38	79.02±0.56
	Hill [44]	✓	42.50±0.16	56.89±0.19	47.31±0.37	49.28±0.09	68.79±0.15	86.70±0.17	78.15±0.99	77.40±0.22
	P-ASL [2]	✓	43.09±0.05	57.67±0.07	53.46±0.22	51.75±0.17	68.95±0.22	87.24±0.13	83.37±0.33	78.96±0.16
PLT-MLC	Head Model (Ours)	✓	47.59±0.09	59.07±0.12	52.35±0.28	53.30±0.19	72.91±0.28	88.59±0.31	82.12±0.27	80.70±0.30
	Tail Model (Ours)	✓	46.30±0.25	58.76±0.29	53.38±0.14	53.09±0.27	71.65±0.34	88.68±0.41	83.51±0.24	80.58±0.36
	COMIC (Ours)	✓	**49.21±0.22**	**60.08±0.13**	**55.36±0.21**	**55.08±0.14**	**73.10±0.35**	**89.18±0.45**	**84.53±0.48**	**81.53±0.35**
Improv. ↑	-	-	1.62～8.16	0.34～3.41	0.53～8.96	1.78～6.16	0.19～6.08	0.29～2.48	-0.3～6.38	0.83～4.13

Table 2: **Ablation study of different modules.** M,C,B represent correction, modification and balance learning, respectively.

Models	Setting			PLT-COCO Dataset		
	M	C	B	*Total mAP*	*Average mAP*	*Recall*
-RLC		✓	✓	54.70±0.13	54.42±0.15	85.26±0.08
-MFM	✓		✓	54.60±0.13	54.33±0.13	84.59±0.19
-HTB	✓	✓		53.65±0.31	53.36±0.31	84.19±0.23
COMIC	✓	✓	✓	**55.08±0.14**	**54.88±0.19**	**88.19±0.22**

4. Experiments

4.1. Experimental Setup

Dataset Construction. The proposed method is analyzed on the created LT versions of two MLC benchmarks (COCO [27] and VOC [8]), called `PLT-COCO` and `PLT-VOC`, respectively. The missing rate of `PLT-COCO` is 40% and it contains 2,962 images from 80 classes. The maximum training number for each class is 1,240 and the minimum number is 6. We select 5000 images from the test set of COCO2017 for evaluation. `PLT-VOC` has the same missing rate setting and contains 2,569 images from 20 classes, in which the maximum training number for each class is 1,117 and the minimum number is 7. We evaluate the performance on VOC2007 test set with 4,952 images. More details about the dataset construction can be found in the Appendix.

Implementation Details. We employ the ResNet-50 [11] as the backbone model to conduct the PLT-MLC task. We train our model with a standard Adam [14] optimizer in all the experiments. The images will be randomly cropped and resized to 224×224 together with standard data augmentation. Besides, we use an identical set of hyperparameters (B=32, Mo=0.9, E_{max}=40)[1] across all the datasets. More details of implementation are in Appendix.

Evaluation Metrics. Following [37], we split these classes into three groups according to the number of their training examples: each *head* class contains over 100 samples as a many shot, each *medium* class has 20 to 100 samples as a medium shot, and each *tail* class has less than 20 samples as a low shot. The total shot indicates all the test samples. We evaluate mean average precision (*mAP*) for all the classes and recall for missing label settings.

Comparison of the Methods. To quantify the efficacy of the proposed framework, we use several baselines for performance comparison according to different aspects[2]. MLC methods: BCE [45], Focal [26], ASL [3]. LT-MLC methods: DB [37], DB-Focal [37] and LWS [13]. PL-MLC methods: Hill [44], Pseudo-Label [15], ML-GCN [5] and P-ASL [2].

4.2. Overall Performance

Table 1 summarizes the quantitative PLT-MLC results of our framework and baselines on `PLT-COCO` and `PLT-VOC`. We make the following observations: 1) In general, irrespective of the different shot scenarios, compared to SoTAs, COMIC achieves the best performance on almost all the metrics across both datasets. In particular, COMIC outperforms other baselines in terms of total shot's mAP by

---

[1]B and Mo refer to the batch size and momentum in the Adam.
[2]We only compare with the approaches that have open source code.

Table 3: **Performance comparison under different missing labeled settings.** 0% indicates an LT dataset that is fully labeled.

Missing Ratio	PLT-COCO Dataset			
	Total Shot	*Many Shot*	*Medium Shot*	*Low Shot*
0%	57.07±0.09	52.21±0.11	59.98±0.12	61.12±0.24
30%	55.80±0.17	49.97±0.11	62.59±0.15	54.56±0.17
40%	54.75±0.19	48.93±0.24	60.31±0.21	54.14±0.21
50%	54.69±0.15	48.74±0.12	56.68±0.16	57.25±0.24
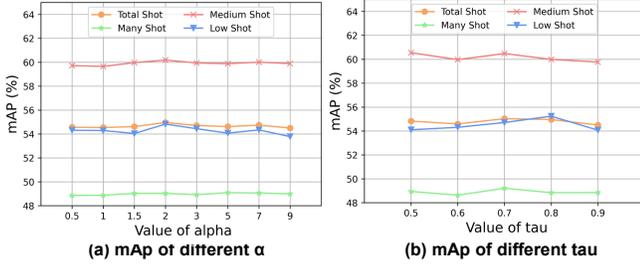
Figure 3: **Ablations with respect to coefficient α and τ.**
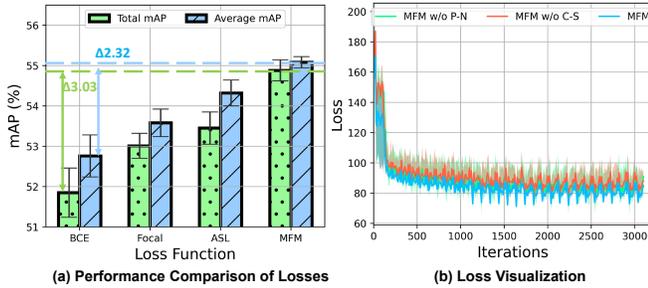
Figure 4: **MLT-MLC results using different losses.**

a large margin (`PLT-COCO`: **1.78% ∼ 6.16%**, `PLT-VOC`: **0.83% ∼ 4.13%**) for PLT-MLC task. 2) Besides, we can observe from Table 1 that the LT methods outperform PL baselines in most $X\text{-}shot$ situations. We believe the underlying reason behind this is that LT data distribution hurts the classification capability for MLC models more seriously than PL. Besides, the label correction may have aggravated the LT issue and further result in performance reduction. 3) Benefiting from the carefully designed HTB module, our COMIC not only achieves the highest total mAP score but also yields balanced results with a narrowed performance gap in different shot metrics. These results demonstrate the superiority of our proposed model.

4.3. Ablation Study

Effectiveness of Each Component. We conduct an ablation study to illustrate the effectiveness of each component in Table 2. Comparing COMIC and COMIC(-RLC) (Row 1 *v.s*
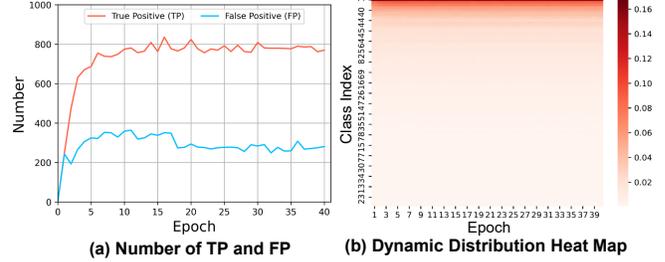

(a) Number of TP and FP
(b) Dynamic Distribution Heat Map

Figure 5: **In-depth analysis of label correction.**

Table 4: **Ablation of MFM.** ↓ indicates the mAP decay.

MFM Factor		PLT-COCO Dataset			
P-N	H-T	*Total Shot*	*Many Shot*	*Medium Shot*	*Low Shot*
	✔	54.44 (↓0.64)	48.65 (↓0.56)	60.00 (↓0.08)	53.81 (↓1.55)
✔		53.70 (↓1.38)	48.38 (↓0.83)	58.99 (↓1.09)	52.91 (↓2.45)
✔	✔	55.08	49.21	60.08	55.36

Row 4), the label (*Correction*) mechanism contributes 0.38% improvement on total mAP. The results of Row 2 show the mAP improvement of the MFM (*Modification*). Meanwhile, Row 3 indicates that it suffers from noticeable performance degradation without the (*Balance*) learning. To sum up, we can observe that the improvement of each module is distinguishable. Combining all the components, our COMIC exhibits steady improvement over the baselines.

Ablation of Missing Rate. To study the effect of partial labels that affect COMIC's results, we evaluate the performance under different missing rates (MR) of labels (from 0% ∼ 50%). Not surprisingly, when the MR decreases, the accuracy of COMIC increases on all the metrics. We also find that the performance gap between different shots is consistently small in all MR settings. The results demonstrate the generalizability of the proposed COMIC that it can produce stable and balanced results under different MR settings.

Hyperparameter α and τ. We investigate the impact of hyper-parameter α and τ for the PLT-MLC task. The mAPs of different hyper-parameter settings on `PLT-COCO` are shown in Figure 3. This figure suggests that the optimal choices of α and τ are around 2 and 0.7, respectively. Either increasing or decreasing these values results in performance decay.

4.4. In-Depth Analysis

We further validate several vital issues of the proposed *Correction → Modification → Balance* learning paradigm by answering the three questions as follows.

Q1: Can the model trust the recalled labels distinguished by RLC? To build the insight on the effectiveness of the label correction mechanism in COMIC, we visualize the *true positive* (TP) and *false positive* (FP) in Figure 5(a)). This figure suggests that the RLC module can distinguish a large
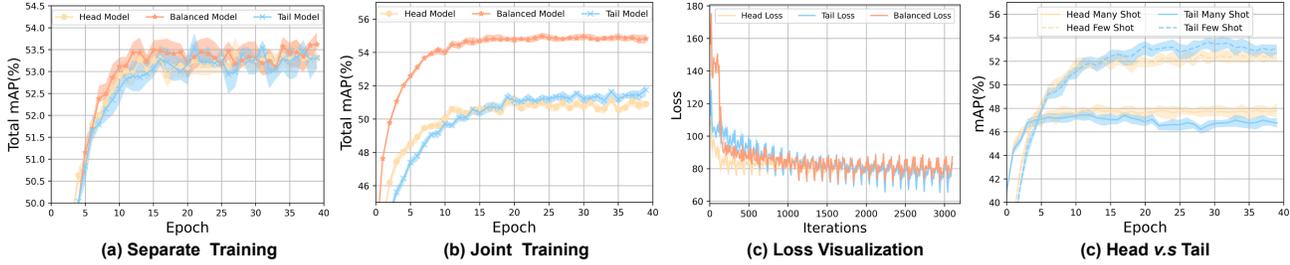
Figure 6: **Analysis of balanced learning of COMIC.** (a) and (b) depict the total mAP of separate and joint training of COMIC within the 40 epochs. (c) summarizes the loss visualization of *head*, balanced and tail models with joint training. (d) demonstrates *head* and *tail* models respectively optimize the *head* and *tail* class's performance.

number of missing labels with high prediction confidence in the early training stage, meanwhile, sum(TP) ≫ sum(FP). However, Figure 5(b) of corrected samples also reveals LT class distribution with respect to the original training set. To address this issue, we dynamically adjust the sample weight conditioned on the real-time distribution to produce a stable performance.

Q2: How does the MFM module affect the PLT-MLC performance? Here, we evaluate the effectiveness of the multi-focal modifier (MFM) loss compared with different loss functions. Figure 4(a) shows the loss ablation results using different losses in our COMIC. Our developed MFM outperforms existing losses, as the designed loss considers the key point of the *head-tail* and *positive-negative* imbalance under the extreme LT distribution in the PLT-MLC task. There are two components in MFM, which are the *positive-negative* (P-N) factor and *head-tail* (H-T) factor. To demonstrate the effect of each component, we train the model with the individual factor in the proposed MFM. As shown in Table. 4, both the P-N factor and the H-T factor play significant roles in MFM. For the H-T factor, it achieves an improvement from 53.7% mAP to 55.08% mAP. Meanwhile, it brings a significant gain on *tail* categories with 2.45% mAP improvement, indicating its effectiveness to alleviate the severe *positive-negative* imbalance problems in the LT class distribution. As for the P-N factor, it brings a steady boost on all shot settings which means it can further alleviate the *positive-negative* issue. Additionally, Figure 4(b) indicates that the MFM loss decreases faster and smoother than the two variants of MFM without different factors, demonstrating its superiority in the PLT-MLC task further.

Q3: How does the HTB module benefit the balanced learning? We systematically present the explicit benefits of the balanced learning strategy in multi-view. 1) Figure 6 (a) and (b) show the comparison between separate and joint training of *head*, balanced and *tail* model with respect to the total mAP on PLT-COCO dataset. An interesting phenomenon is that the *detached head* and *tail* models slightly outperform the *joint head* and *tail* models but suffer

from an unstable performance. In contrast, the accuracy of the *joint* trained balanced model increases much faster and smoother than the *detached* balanced model which also yields a stable performance and faster convergence speed. This phenomenon is reasonable as the main optimization objective in joint training is to improve the balanced model's performance. It can be regarded as the knowledge distillation effect that enables the balanced model to learn from the *head biased* and *tail biased* model, and this in turn facilitates the PLT-MLC learning. 2) During the competition of *head v.s tail*, the *head* model's loss drops faster (shown in Figure 6 (c)) and is biased to optimizing the *head* samples, while the *tail* model produces an opposite result. Such *head* and *tail* biased results form a foundation that enables our COMIC to be trained in a balanced and stable learning effect. 3) We also perform the analysis of the different balanced learning blocks in our COMIC. As presented in Table 6 (in the Appendix), the DL, NC and AMV contribute 0.05%, 1.08% and 0.37% improvement on total shot mAP. The observations and analysis verify the effectiveness of balanced learning for being able to study from the *head* and *tail* models, thereby achieving the PLT-MLC improvement.

5. Conclusions

We have presented a fire-new task called PLT-MLC and correspondingly developed a novel framework, named COMIC. COMIC simultaneously addresses the partial labeling and long-tailed environments in a *Correction → Modification → Balance* learning manner. On two newly proposed benchmarks, PLT-COCO and PLT-VOC, we demonstrate that the proposed framework significantly outperforms existing MLC, LT-MLC and PL-MLC approaches.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] Emanuel Ben-Baruch, Tal Ridnik, Itamar Friedman, Avi Ben-Cohen, Nadav Zamir, Asaf Noy, and Lihi Zelnik-Manor. Multi-label classification with partial annotations using class-aware selective loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4764–4772, 2022.

[3] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. *arXiv preprint arXiv:2009.14119*, 2020.

[4] Serhat Selcuk Bucak, Rong Jin, and Anil K Jain. Multi-label learning with incomplete class assignments. In *CVPR 2011*, pages 2801–2808. IEEE, 2011.

[5] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5177–5186, 2019.

[6] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *Advances in neural information processing systems*, 31, 2018.

[7] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 647–657, 2019.

[8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[9] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4367–4375, 2018.

[10] Hao Guo and Song Wang. Long-tailed multi-label visual recognition by collaborative training on uniform and re-balanced samplings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15089–15098, 2021.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, pages 67–84. Springer, 2016.

[13] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013.

[16] Juncheng Li, Minghe Gao, Longhui Wei, Siliang Tang, Wenqiao Zhang, Mengze Li, Wei Ji, Qi Tian, Tat-Seng Chua, and Yueting Zhuang. Gradient-regulated meta-prompt learning for generalizable vision-language models. 2023.

[17] Juncheng Li, XIN HE, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. Fine-grained semantically aligned vision-language pre-training. In *Advances in Neural Information Processing Systems*, 2022.

[18] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, and Yueting Zhuang. Empowering vision-language models to follow interleaved vision-language instructions, 2023.

[19] Juncheng Li, Siliang Tang, Linchao Zhu, Wenqiao Zhang, Yi Yang, Tat-Seng Chua, and Fei Wu. Variational cross-graph reasoning and adaptive structured semantics learning for compositional temporal grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[20] Juncheng Li, Xin Wang, Siliang Tang, Haizhou Shi, Fei Wu, Yueting Zhuang, and William Yang Wang. Unsupervised reinforcement learning of transferable meta-skills for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12123–12132, 2020.

[21] Li-Jia Li and Li Fei-Fei. Optimol: automatic online picture collection via incremental model learning. *International journal of computer vision*, 88(2):147–168, 2010.

[22] Mengze Li, Han Wang, Wenqiao Zhang, Jiaxu Miao, Zhou Zhao, Shengyu Zhang, Wei Ji, and Fei Wu. Winner: Weakly-supervised hierarchical decomposition and alignment for spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23090–23099, 2023.

[23] Mengze Li, Tianbao Wang, Jiahe Xu, Kairong Han, Shengyu Zhang, Zhou Zhao, Jiaxu Miao, Wenqiao Zhang, Shiliang Pu, and Fei Wu. Multi-modal action chain abductive reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4617–4628, 2023.

[24] Mengze Li, Tianbao Wang, Haoyu Zhang, Shengyu Zhang, Zhou Zhao, Jiaxu Miao, Wenqiao Zhang, Wenming Tan, Jin Wang, Peng Wang, et al. End-to-end modeling via information tree for one-shot natural language spatial video grounding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8707–8717, 2022.

[25] Mengze Li, Tianbao Wang, Haoyu Zhang, Shengyu Zhang, Zhou Zhao, Wenqiao Zhang, Jiaxu Miao, Shiliang Pu, and Fei Wu. Hero: Hierarchical spatio-temporal reasoning with contrastive action correspondence for end-to-end video object grounding. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3801–3810, 2022.

[26] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[28] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019.

[29] SGOPAL Patro and Kishore Kumar Sahu. Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*, 2015.

[30] William J Reed. The pareto, zipf and other power laws. *Economics letters*, 74(1):15–19, 2001.

[31] Yu-Yin Sun, Yin Zhang, and Zhi-Hua Zhou. Multi-label learning with weak label. In *Twenty-fourth AAAI conference on artificial intelligence*, 2010.

[32] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013.

[33] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in Neural Information Processing Systems*, 33:1513–1524, 2020.

[34] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[36] Jonatas Wehrmann, Ricardo Cerri, and Rodrigo Barros. Hierarchical multi-label classification networks. In *International conference on machine learning*, pages 5075–5084. PMLR, 2018.

[37] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *European Conference on Computer Vision*, pages 162–178. Springer, 2020.

[38] Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. Sgm: sequence generation model for multi-label classification. *arXiv preprint arXiv:1806.04822*, 2018.

[39] Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit Dhillon. Large-scale multi-label learning with missing labels. In *International conference on machine learning*, pages 593–601. PMLR, 2014.

[40] Wenqiao Zhang, Changshuo Liu, Can Cui, and Beng Chin Ooi. Causal and collaborative proxy-tasks learning for semi-supervised domain adaptation. *arXiv preprint arXiv:2303.17526*, 2023.

[41] Wenqiao Zhang, Haochen Shi, Jiannan Guo, Shengyu Zhang, Qingpeng Cai, Juncheng Li, Sihui Luo, and Yueting Zhuang. Magic: Multimodal relational graph adversarial inference for diverse and unpaired text-based image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3335–3343, 2022.

[42] Wenqiao Zhang, Siliang Tang, Yanpeng Cao, Shiliang Pu, Fei Wu, and Yueting Zhuang. Frame augmented alternating attention network for video question answering. *IEEE Transactions on Multimedia*, 22(4):1032–1041, 2019.

[43] Wenqiao Zhang, Lei Zhu, James Hallinan, Shengyu Zhang, Andrew Makmur, Qingpeng Cai, and Beng Chin Ooi. Boostmis: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20666–20676, 2022.

[44] 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.

[45] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

[46] Pengfei Zhu, Qian Xu, Qinghua Hu, Changqing Zhang, and Hong Zhao. Multi-label feature selection with missing labels. *Pattern Recognition*, 74:488–502, 2018.