

Positive Label Is All You Need for Multi-Label Classification

Zhixiang Yuan*
Anhui University of Technology
Maanshan, China
zxyuan@ahut.edu.cn

Kaixin Zhang*†
Anhui University of Technology
Maanshan, China
kxzhang0618@163.com

Tao Huang
The University of Sydney
Darlington, Australia
thua7590@uni.sydney.edu.au

Abstract—Multi-label classification (MLC) faces challenges from label noise in training data due to annotating diverse semantic labels for each image. Current methods mainly target identifying and correcting label mistakes using trained MLC models, but still struggle with persistent noisy labels during training, resulting in imprecise recognition and reduced performance. Our paper addresses label noise in MLC by introducing a positive and unlabeled multi-label classification (PU-MLC) method. To counteract noisy labels, we directly discard negative labels, focusing on the abundance of negative labels and the origin of most noisy labels. PU-MLC employs positive-unlabeled learning, training the model with only positive labels and unlabeled data. The method incorporates adaptive re-balance factors and temperature coefficients in the loss function to address label distribution imbalance and prevent over-smoothing of probabilities during training. Additionally, we introduce a local-global convolution module to capture both local and global dependencies in the image without requiring backbone retraining. PU-MLC proves effective on MLC and MLC with partial labels (MLC-PL) tasks, demonstrating significant improvements on MSCOCO and PASCAL VOC datasets with fewer annotations. Code is available at: <https://github.com/TAKELAMAG/PU-MLC>.

Index Terms—Multi-label classification, image recognition, positive-unlabeled learning, noisy label

I. INTRODUCTION

Recently, multi-label classification (MLC) [1]–[3] has gained significant attention as a natural image often contains multiple objects or concepts. Traditional approaches to MLC treat it as a series of binary classification tasks, each determining the presence or absence of individual classes.

Noisy labels, a prevalent issue in MLC datasets due to annotation difficulties [3], disrupt training and impair performance (see Figure 1 (a)-(b)). To address this, certain methods [3] suggest initially training models with these noisy labels, then using the trained model to correct or eliminate mislabeled data. However, the involvement of mislabeled labels in training phase can still negatively influence the process and potentially lead to inaccuracies in identifying noisy labels.

The mislabeling issue becomes more pronounced in the context of multi-label classification with partial labels (MLC-PL) [4]–[6]. In MLC-PL, where models are trained with partially labeled datasets to minimize annotation costs (refer to Figure 1 (c)), the limited label information increases the model’s vulnerability to label noise. Addressing this,

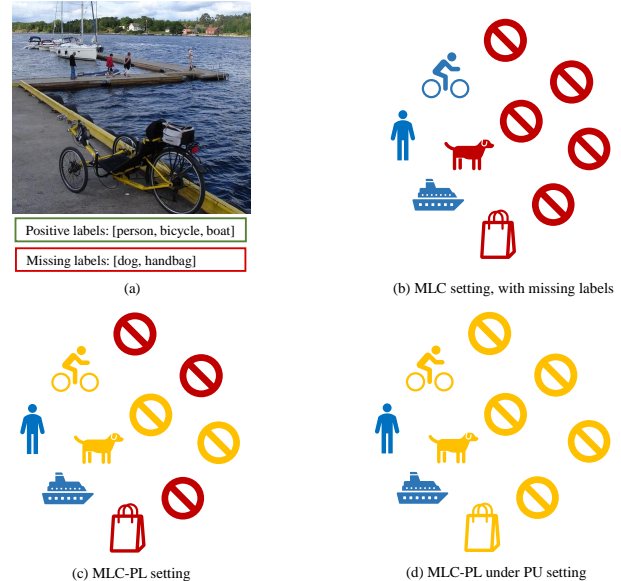


Fig. 1. Comparisons of different learning methods in MLC. (a) an image which has two missing labels. To train the sample image, (b) missing labels in traditional MLC methods are mistakenly classified as negative labels; (c) MLC-PL samples a proportion of labels, but still encounters false negative labels; (d) Our method treats all negative labels as unlabeled ones. Blue, red, and yellow icons denote positive, negative, and unknown labels, respectively.

some approaches [4] aim to mitigate noisy labels’ impact by adjusting the loss weight for each sample. Others explore semantic-aware representations for pseudo label generation [5] or blend category-specific semantic representations across different images [6]. However, these MLC-PL methods, like their MLC counterparts, still incorporate mislabeled samples in training. This practice can lead to inaccurate loss weight assessments and pseudo label creation, ultimately impacting model performance.

To address the issue of noisy labels in multi-label classification (MLC) and MLC with partial labels (MLC-PL), which impair training performance, we propose a novel approach in the absence of a reliable method to identify these noisy labels: removing all labels. Drawing inspiration from positive-unlabeled (PU) learning [7], [8], which trains classifiers using only positive labels and compares favorably with traditional positive-negative (PN) learning (refer to Figure 1(d)), our

*Equal contributions. †Corresponding author.

method discards all negative labels and relies on positive and unlabeled data for training MLC models. This strategy, leveraging the imbalance of negative labels in MLC datasets (see Figure 2(b)), reduces annotation errors. PU learning, known for its robustness and accuracy, especially with noisy negative labels, uses an unbiased risk estimator for better performance. It provides more accurate and informative labeling using soft labels, contrasting with hard labels in conventional methods.

As a result, we introduce a novel method, positive and unlabeled multi-label classification (PU-MLC), adapting PU learning for MLC tasks by integrating multiple binary classifications. To address the significant imbalance between positive and negative labels in MLC, we introduce an adaptive rebalance factor in the PU loss to adjust loss weights effectively. Recognizing the complexity of training multiple binary tasks in MLC compared to standard PU learning, we propose an adaptive temperature coefficient module. This module fine-tunes the sharpness of predicted probabilities in the loss function, preventing over-smoothing in early training stages and enhancing optimization. Additionally, we present a novel local-global convolution module that incorporates both local and global image dependencies. This module enriches existing convolution layers with global information without requiring backbone retraining.

Our PU-MLC method is both simple and effective for MLC and PU-MLC tasks. It demonstrates strong performance even with limited positive labels, reducing annotation costs. Our extensive experiments on benchmark datasets MS-COCO [9] and PASCAL VOC 2007 [10] show that PU-MLC significantly improves performance in both MLC and MLC-PL settings, while utilizing fewer annotated labels.

II. RELATED WORK

A. Multi-Label Classification

Multi-label classification (MLC) task aims to recognize semantic categories in a given image, which usually contains multiple objects or concepts. Previous works [1], [11], [12] propose to construct pairwise statistical correlations using the first-order adjacency matrix obtained by graph convolutional networks (GCN) [13]. Although the above methods achieve noteworthy success, they cannot extract higher-order correlations and can attract overfitting on small training sets. Some works [14], [15] introduce transformer to extract complicated dependencies among visual features and labels.

MLC with partial labels (MLC-PL). Traditional multi-label classification (MLC) tasks rely on fully annotated datasets, and making such datasets is expensive, time-consuming, and error-prone. To reduce the cost of annotation, multi-label classification with partial labels (MLC-PL) attempts to train models with partially-annotated labels per image, which both contain positive and negative labels. Recent works [4], [5], [16] propose to generate pseudo labels to those unknown samples based on the learned knowledge in the training model, and then train the model with ground-truth partial labels and generated pseudo labels.

B. Positive-Unlabeled (PU) learning

Different from the traditional positive-negative (PN) learning in the binary classification task, PU learning aims to train the model with only positive and unknown labels [17]. Recent advances [7], [8], [18], [19] have achieved remarkable progress in deep learning. However, these methods rely heavily on the class prior estimation. While the class prior in the training dataset may not always correctly represent the label distribution in the validation set, and thus performing PU learning without class prior becomes an emergent topic [8], [20]–[22]. For example, vPU [8] proposes a variational principle to achieve superior performance without class prior. In this paper, we extend PU learning to MLC task based on vPU [8].

III. PROPOSED APPROACH: PU-MLC

A. MLC as PU learning

MLC as PN learning. MLC task is usually formulated as multiple binary classification sub-tasks, and each sub-task aims to recognize whether a specific category is in the input image. Formally, for a MLC task with C categories, let $\mathbf{s} \in \mathbb{R}^{N \times C}$ and $\mathbf{y} \in \{-1, +1\}^{N \times C}$ be the predicted logits and the ground-truth positive and negative (PN) labels, respectively, where N denotes batch size, the overall classification loss is formulated as

$$\mathcal{L}_{\text{mlc}} = \frac{1}{C \times N} \sum_{c=1}^C \sum_{n=1}^N [\mathbb{1}(y_{n,c} = +1) \mathcal{L}_+(\sigma(s_{n,c})) + \mathbb{1}(y_{n,c} = -1) \mathcal{L}_-(\sigma(s_{n,c}))], \quad (1)$$

where $\sigma(\cdot)$ is the Sigmoid function, $\mathbb{1}(\cdot)$ is an indicator function that takes the value 1 only if the condition is true and 0 otherwise, \mathcal{L}_+ and \mathcal{L}_- denote losses on positive and negative labels, respectively.

Before presenting our PU-learning based MLC method, we first rewrite the learning objective of the above positive-negative (PN) classification loss ((1)) as the expected risk on the training set. The total risk R_{mlc} is accumulated with all PN sub-tasks, and for each task (category) with the class prior (proportion of positive labels) π_p and $\mathbf{S} \in \mathbb{R}^M$ being its corresponding logits on the training set with M images, its risk is formulated as

$$R_{\text{pn}} = \pi_p \mathbb{E}_{\mathcal{P}}[\mathcal{L}_+(\sigma(\mathbf{S}))] + (1 - \pi_p) \mathbb{E}_{\mathcal{N}}[\mathcal{L}_-(\sigma(\mathbf{S}))], \quad (2)$$

where the images regarding to their label types are split into positive set \mathcal{P} and negative set \mathcal{N} , and we have the expectations of positive and negative losses

$$\begin{aligned} \mathbb{E}_{\mathcal{P}}[\mathcal{L}_+(\sigma(\mathbf{S}))] &= \frac{1}{|\mathcal{P}|} \sum_{s_m \in \mathcal{P}} \mathcal{L}_+(\sigma(s_m)), \\ \mathbb{E}_{\mathcal{N}}[\mathcal{L}_-(\sigma(\mathbf{S}))] &= \frac{1}{|\mathcal{N}|} \sum_{s_m \in \mathcal{N}} \mathcal{L}_-(\sigma(s_m)). \end{aligned} \quad (3)$$

PN to PU. In this paper, we aim to train a MLC model with only positive labels; *i.e.*, our training set is composed of a positive set \mathcal{P} and an unlabeled set \mathcal{U} (mixture of unlabeled positive and negative images). Nevertheless, the negative labels

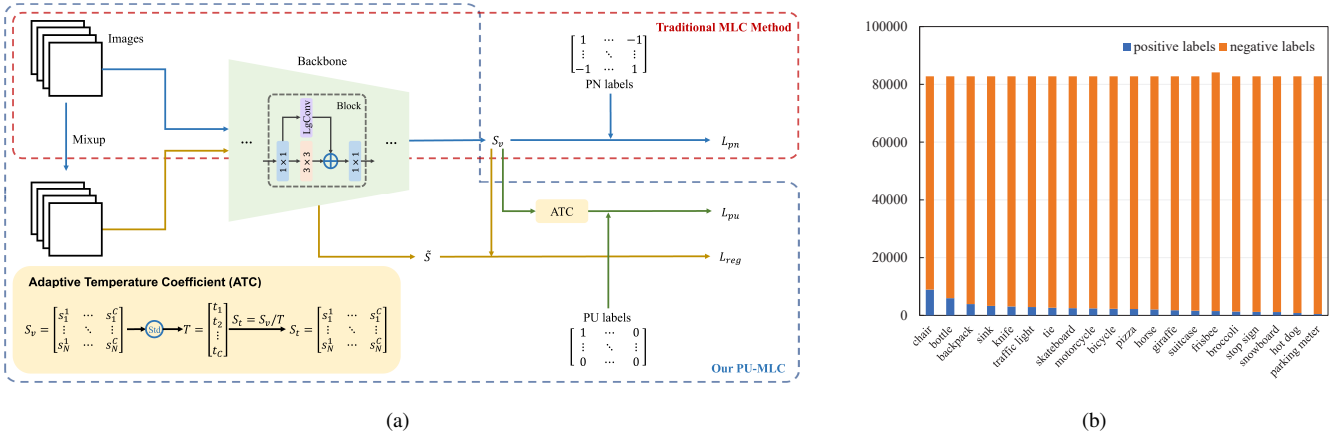


Fig. 2. (a) Overview of our proposed PU-MLC. Instead of using positive and negative labels in the traditional MLC method (red box), our PU-MLC conducts a positive-unlabeled (PU) learning strategy with only partial positive labels leveraged. Besides, we introduce mixup regularization loss and the adaptive temperature coefficient module to further boost the performance. Additionally, we enhance the global representations in backbone by integrating a local-global convolution module to every 3×3 local convolutions. *Std*: standard deviation. (b) Histograms of the number of positive and negative labels in each category. We randomly select 20 categories from MS-COCO train set.

are unavailable in our PU setting, and therefore we cannot directly optimize (2) to obtain our model. In order to train a classifier with positive and unknown labels, a classical method uPU [7] introduces an unbiased formulation to the PN learning by rewriting the expectation of negative classification loss $\mathbb{E}_{\mathcal{N}}[\mathcal{L}_-(\sigma(\mathbf{S}))]$ to

$$(1 - \pi_p)\mathbb{E}_{\mathcal{N}}[\mathcal{L}_-(\sigma(\mathbf{S}))] = \mathbb{E}_{\mathcal{U}}[\mathcal{L}_-(\sigma(\mathbf{S}))] - \pi_p\mathbb{E}_{\mathcal{P}}[\mathcal{L}_-(\sigma(\mathbf{S}))], \quad (4)$$

and thus (2) could be converted to PU format:

$$R_{\text{pu}} = \pi_p\mathbb{E}_{\mathcal{P}}[\mathcal{L}_+(\sigma(\mathbf{S}))] - \pi_p\mathbb{E}_{\mathcal{P}}[\mathcal{L}_-(\sigma(\mathbf{S}))] + \mathbb{E}_{\mathcal{U}}[\mathcal{L}_-(\sigma(\mathbf{S}))], \quad (5)$$

However, the above method easily causes overfitting in deep neural networks and rely heavily on the class prior, and we empirically find that it performs poorly on the multi-label classification task, as the task is more challenging and many categories have very small class priors. Hence, this paper utilizes a recent PU learning method vPU [8], which proposes a novel loss function based on the variational principle to approximate the ideal classifier without the class prior:

$$R_{\text{var}} = \log \mathbb{E}_{\mathcal{U}}[\sigma(\mathbf{S})] - \mathbb{E}_{\mathcal{P}}[\log \sigma(\mathbf{S})]. \quad (6)$$

Hence, for each category c , the classification loss becomes

$$\mathcal{L}_{\text{var}}^{(c)} = \log\left(\frac{1}{|\mathcal{U}_N^{(c)}|} \sum_{s_u \in \mathcal{U}_N^{(c)}} \sigma(s_u)\right) - \frac{1}{|\mathcal{P}_N^{(c)}|} \sum_{s_p \in \mathcal{P}_N^{(c)}} \log \sigma(s_p), \quad (7)$$

here $\mathcal{P}_N^{(c)}$ and $\mathcal{U}_N^{(c)}$ denote positive samples and unlabeled samples of category c in each mini-batch, respectively. Note that vPU also introduces a consistency regularization term $\mathcal{L}_{\text{reg}}^{(c)}$ based on Mixup [23], which alleviates the overfitting problem and increases the robustness in PU learning.

As a result, in our PU-MLC, the traditional MLC loss in (1) is replaced with our PU loss, and the overall loss function is formulated as

$$\mathcal{L}_{\text{pu-mlc}} = \sum_{c=1}^C (\mathcal{L}_{\text{var}}^{(c)} + \lambda \mathcal{L}_{\text{reg}}^{(c)}), \quad (8)$$

where λ is a scalar to balance the losses and we set $\lambda = 1$ in all experiments.

Importantly, our approach diverges from traditional PU learning by including all positive samples \mathcal{P} into \mathcal{U} . This ensures that \mathcal{U} maintains a label distribution similar to a conventional training set, a critical factor for the effectiveness of PU learning (refer to our ablation studies for further details).

B. Catastrophic Imbalance of Label Distribution

MLC datasets typically have a far greater number of negative than positive labels, as shown in Figure 2(b). In PU-MLC, where all negative labels are moved to the unlabeled set, there is a significant imbalance in the number of samples affecting the two terms of \mathcal{L}_{var} in equation (7) within each mini-batch. This differs from conventional PU learning where positive and negative samples are equal in batch size. Applying (7) as is in our method would cause the unlabeled term to overly influence the optimization, leading to suboptimal results in MLC-PL, especially at low known label ratios (e.g., only achieving 51.8% mAP with 10% positive labels).

To alleviate the catastrophic imbalance of label distribution, we aim to narrow down the loss weight of unlabeled term to decrease its importance in optimization. Inspired by focal loss [24] and ASL [3], we propose a re-balance factor to dynamically re-weight the unlabeled loss based on the predicted probabilities on unlabeled samples, and (7) is reformulated as

$$\mathcal{L}_{\text{var}}^{(c)} = p_c^\gamma \log\left(\frac{1}{|\mathcal{U}_N^{(c)}|} \sum_{s_u \in \mathcal{U}_N^{(c)}} \sigma(s_u)\right) - \frac{1}{|\mathcal{P}_N^{(c)}|} \sum_{s_p \in \mathcal{P}_N^{(c)}} \log \sigma(s_p), \quad (9)$$

where p_c^γ denotes our re-balance factor, with $p_c = \frac{1}{|\mathcal{U}|} \sum_{s_u \in \mathcal{U}} \sigma(s_u)$ being the mean probability of unlabeled samples, and γ is used to control the value of the factor. In our experiments, we set larger γ for smaller known label ratios, as the imbalance is severer on smaller ratios and we need a smaller weight on unlabeled loss to balance the loss.

C. Adaptive Temperature Coefficient

In PU learning, the model serves as an estimator for probabilistic evaluations of unlabeled samples, optimizing them via the unlabeled loss term [17]. However, the task in MLC, which involves learning multiple binary classifiers, is considerably more complex than the single binary classification task in standard PU methods. This complexity results in a slower convergence rate during the early stages of training. Consequently, the predicted probability distribution tends to be over-smooth, reducing the effectiveness of the unlabeled loss.

To adjust the smoothness of probabilistic distribution, we follow [25] and propose a temperature coefficient τ to scale the logit values, *i.e.*, $s_t = s/\tau$, then the s_t is fed into the PU loss in place of the original s .

By setting $\tau < 1$, the probabilistic distribution becomes sharper, providing more meaningful and impactful feedback to the loss function. However, our empirical findings indicate that a fixed temperature coefficient τ enhances performance only under certain known label ratios and specific datasets (refer to Table 3 in the appendix). For instance, the MS-COCO dataset benefits from $\tau < 1$, whereas the PASCAL VOC dataset shows better results with $\tau > 1$. This suggests that the optimal τ varies not only across different datasets but also among different categories within the same dataset, necessitating individual adjustments.

As a result, we propose an adaptive temperature coefficient module to first measure the sharpness of each category in each batch, then apply independent temperatures on each category. Formally, given the predicted logits s , the sharpness of each category c is measured using the standard deviation of the logits, and then the temperature is obtained by multiplying a scalar α onto the sharpness value, *i.e.*,

$$\tau^{(c)} = \min(\alpha \cdot \text{Std}(s_c), 1). \quad (10)$$

We use a minimum function to ensure that the $\tau^{(c)}$ is less than or equal to 1, since we do not want the $\tau^{(c)}$ to exceed 1, which could even exacerbate the over-smooth.

The final PU loss $\mathcal{L}_{\text{val}}^{(c)}$ becomes

$$\begin{aligned} \mathcal{L}_{\text{var}}^{(c)} = & p_c^\gamma \log\left(\frac{1}{|\mathcal{U}_N^{(c)}|} \sum_{s_u \in \mathcal{U}_N^{(c)}} \sigma(s_u/\tau^{(c)})\right) \\ & - \frac{1}{|\mathcal{P}_N^{(c)}|} \sum_{s_p \in \mathcal{P}_N^{(c)}} \log \sigma(s_p/\tau^{(c)}). \end{aligned} \quad (11)$$

Our adaptive temperature coefficient is suitable for different known label ratios and datasets, which could gain consistent improvements. The overall framework of our model is illustrated in Figure 2(a).

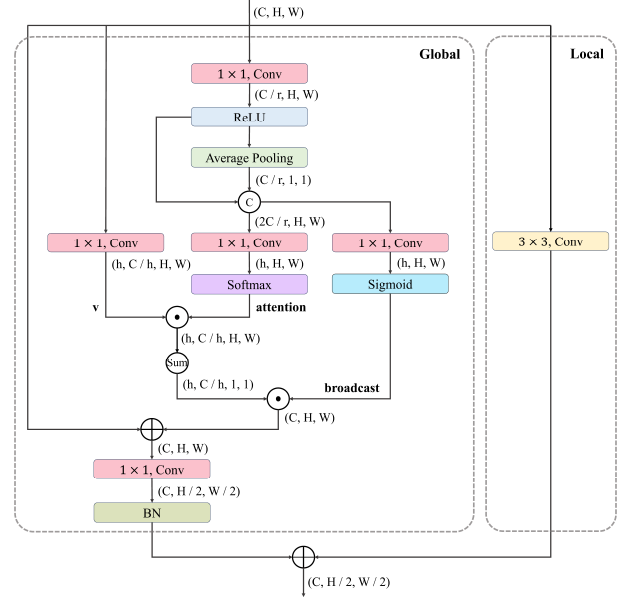


Fig. 3. Illustration of Local-global convolution.

D. Local-Global Convolution

Vision transformers [28], [29] have shown notable advancements over classical CNNs by capturing global dependencies, yet they face challenges like high memory usage, deployment difficulties, and limitations in lightweight models. Addressing these, we introduce a convolution-based global module, LgConv, designed as a plug-and-play enhancement for CNNs without necessitating retraining of the backbone.

As illustrated in Figure 3, LgConv augments traditional local convolution with a global branch. This branch first transforms input features to incorporate both local and global information (via average pooling), followed by two 1×1 convolutions creating spatial multi-head attentions and a broadcast attention. Softmax and Sigmoid functions are then applied. The process concludes with a 1×1 convolution and batch normalization to project the feature.

To preserve the pretrained backbone’s semantic integrity, we initialize the global branch’s scale parameters γ in the final batch normalization layer at a minimal value (0.0001). This ensures the global branch’s initial influence on original features is minimal, allowing for a smooth evolution of the backbone during training.

IV. EXPERIMENTS

To verify the efficacy of PU-MLC, we conduct extensive experiments on two popular benchmarks MS-COCO [9] and PASCAL VOC [10]. We adopt similar training strategies following previous works [5], [6], which will be detailedly discussed in appendix.

A. Results on MS-COCO

MLC-PL setting. To demonstrate the effectiveness of the PU-MLC, we compare our PU-MLC with current published

TABLE I

THE COMPARISONS ON MS-COCO AND VOC 2007 UNDER DIFFERENT KNOWN LABEL RATIOS. NOTE THAT OUR PU-MLC ONLY USES PARTIAL POSITIVE LABELS, WHILE OTHER METHODS TRAIN MODELS WITH THE SAME NUMBER OF POSITIVE LABELS AND ADDITIONAL NEGATIVE LABELS. * INDICATES THE BACKBONE IS PRETRAINED BY CLIP [26]. RESULTS EXCEPT DUALCOOP AND OUR METHOD ARE REPORTED BY SARB [6].

Datasets	Methods	10%	20%	30%	40%	50%	60%	70%	80%	90%	Avg. mAP	Avg. OF1	Avg. CF1
MS-COCO	ASL [3]	69.7	74.0	75.1	76.8	77.5	78.1	78.7	79.1	79.7	76.5	46.7	47.9
	CL [4]	26.7	31.8	51.5	65.4	70.0	71.9	74.0	77.4	78.0	60.7	61.9	48.3
	Partial BCE [4]	61.6	70.5	74.1	76.3	77.2	77.7	78.2	78.4	78.5	74.7	74.0	68.8
	SST [5]	68.1	73.5	75.9	77.3	78.1	78.9	79.2	79.6	79.9	76.7	-	-
	SARB [6]	72.5	76.0	77.6	78.7	79.6	79.8	80.0	80.5	80.8	78.4	76.8	72.7
	PU-MLC	75.7	78.6	80.2	81.3	82.0	82.6	83.0	83.5	83.8	81.2	77.4	75.7
	DualCoOp* [27]	78.7	80.9	81.7	82.0	82.5	82.7	82.8	83.0	83.1	81.9	78.1	75.3
PU-MLC*	80.2	83.2	84.4	85.6	85.9	86.6	87.0	87.1	87.5	85.3	81.7	79.1	
VOC 2007	ASL [3]	82.9	88.6	90.0	91.2	91.7	92.2	92.4	92.5	92.6	90.5	41.0	40.9
	CL [4]	44.7	76.8	88.6	90.2	90.7	91.1	91.6	91.7	91.9	84.1	83.8	75.4
	Partial BCE [4]	80.7	88.4	89.9	90.7	91.2	91.8	92.3	92.4	92.5	90.0	87.9	84.8
	SST [5]	81.5	89.0	90.3	91.0	91.6	92.0	92.5	92.6	92.7	90.4	-	-
	SARB [6]	85.7	89.8	91.8	92.0	92.3	92.7	92.9	93.1	93.2	91.5	88.3	86.0
	PU-MLC	88.0	90.7	91.9	92.0	92.4	92.7	93.0	93.4	93.5	92.0	88.2	86.5
	DualCoOp* [27]	90.3	92.2	92.8	93.3	93.6	93.9	94.0	94.1	94.2	93.2	86.3	84.2
PU-MLC*	91.3	92.9	93.3	93.7	93.8	94.3	94.5	94.6	94.8	93.7	89.8	88.2	

TABLE II

COMPARISONS OF THE NUMBER OF ANNOTATED LABELS USED IN TRAINING ON MS-COCO. *Reduction*: THE REDUCTION RATIO ON USED TRAINING ANNOTATIONS OF OUR METHOD COMPARED TO OTHERS.

Methods	PU-MLC			Others		
	10%	50%	100%	10%	50%	100%
Positive	24,103	120,517	241,035	24,103	120,517	241,035
Negative	0	0	0	638,160	3,190,802	6,381,605
Total	24,103	120,517	241,035	662,263	3,311,319	6,622,640
Reduction	-96.4%	-96.6%	-96.4%	-	-	-

state-of-the-art methods. As the experimental results shown in Table I, our PU-MLC significantly outperforms previous methods under different known label ratios. For example, on a high known label ratio of 90%, we obviously surpass SARB by 3.0% in mAP. Compared with previous methods, our method achieves state-of-the-art results in average mAP, OF1 and CF1, which are 81.2%, 77.4% and 75.7%, respectively. DualCoOp uses CLIP [26], a large-scale vision-language pre-trained model, as its backbone to achieve exceptional performance. For a fair comparison, by only using the same visual model, our method achieves superior performance than DualCoOp with both visual and language models.

Note that these significant improvements are obtained with even fewer annotated labels used in training compared to other methods (*e.g.*, with 10% known label ratio, we only use 10% positive labels, while other methods use 10% positive labels and 10% negative labels), this indicates that our method is more effective and efficient on limited training annotations. As shown in Table II, the number of annotated labels used by PU-MLC in model training is much smaller than other methods based on PN. Concretely, our method achieves the best results while decreasing the amount of annotated labels by 96.4% at each known label ratio.

MLC setting. Since our method is designed for both

TABLE III

MAP ON MS-COCO IN MLC SETTING.

Methods	mAP	OF1	CF1
ResNet-101 [32]	77.3	76.8	72.8
Cop [33]	81.1	75.1	72.7
CADM [2]	82.3	79.6	77.0
ML-GCN [1]	83.0	80.3	78.0
PU-MLC	84.2	79.1	78.2

MLC and MLC-PL tasks, we also conduct experiments to validate our performance on traditional MLC. As shown in Table III, we achieve promising performance compared to previous methods. Similar to MLC-PL, our method in MLC is trained with only positive labels, and discards a large number of negative labels (negative labels are $\sim 26.5\times$ more than positive labels), our results can still outperform those methods trained with full annotations. Besides, compared with our PN learning baseline ResNet-101, our MLC-PL significantly outperforms it by 6.9% in mAP, which demonstrates that our method is beneficial to MLC setting by ignoring those noisy negative labels.

B. Results on Pascal VOC 2007

Table I shows the comparisons between PU-MLC and state-of-the-art methods on Pascal VOC. Although Pascal VOC has a small size of the sample and simple categories, and many previous methods achieve splendid results, we still outperform them on average mAP and CF1. Especially on the most challenging 10% known labels, we obviously surpass SARB by 2.3% in mAP. On high known label ratios, our improvements are not as significant as that in MS-COCO dataset, a possible reason is that VOC dataset is much easier and smaller than MS-COCO, and using the previous methods can also obtain impressive performance. Additionally, we compare our method with DualCoOp. By using only the same visual model, our

approach achieves improvements across all the known label ratios.

V. CONCLUSION

In this paper, we propose positive and unlabeled multi-label classification (PU-MLC). By removing all the negative labels in training, our method benefits from the cleaner annotations. Besides, we introduce an adaptive re-balance factor and adaptive temperature coefficient to better adapt PU learning in MLC task, which achieves significant improvements, especially on small known label proportions. Finally, we design a local-global convolution module to effectively capture both local and global dependencies within the image. Extensive experiments on MS-COCO and PASCAL VOC datasets demonstrate our efficacy. Adopting more advanced PU learning methods and combining recent approaches on model architectures in MLC would be a potential direction of improving PU-MLC.

REFERENCES

- [1] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5177–5186.
- [2] Z.-M. Chen, X.-S. Wei, X. Jin, , and Y. Guo, "Multi-label image recognition with joint class-aware map disentangling and label correlation embedding," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 2019, pp. 622–627.
- [3] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Asymmetric loss for multi-label classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 82–91.
- [4] D. Huynh and E. Elhamifar, "Interactive multi-label cnn learning with partial labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9423–9432.
- [5] T. Chen, T. Pu, H. Wu, Y. Xie, and L. Lin, "Structured semantic transfer for multi-label recognition with partial labels," in *Proceedings of the AAAI conference on artificial intelligence*, 2022, pp. 339–346.
- [6] T. Pu, T. Chen, H. Wu, Y. Lu, and L. Lin, "Semantic-aware representation blending for multi-label image recognition with partial labels," 2022.
- [7] M. D. Plessis, G. Niu, and M. Sugiyama, "Convex formulation for learning from positive and unlabeled data," in *International conference on machine learning*, 2015, pp. 1386–1394.
- [8] H. Chen, F. Liu, Y. Wang, L. Zhao, and H. Wu, "A variational approach for learning from positive and unlabeled data," in *Advances in Neural Information Processing Systems*, 2020, pp. 14 844–14 854.
- [9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014, pp. 740–755.
- [10] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.
- [11] T. Chen, M. Xu, X. Hui, H. Wu, and L. Lin, "Learning semantic-specific graph representation for multi-label image recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 522–531.
- [12] R. You, Z. Guo, L. Cui, X. Long, Y. Bao, and S. Wen, "Cross-modality attention with semantic graph embedding for multi-label classification," *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 7, pp. 12 709–12 716, 2020.
- [13] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016.
- [14] J. Lanchantin, T. Wang, V. Ordonez, and Y. Qi, "General multi-label image classification with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 16 478–16 488.
- [15] J. Zhao, K. Yan, Y. Zhao, X. Guo, F. Huang, and J. Li, "Transformer-based dual relation graph for multi-label image recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 163–172.
- [16] T. Durand, N. Mehrasa, and G. 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 (CVPR)*, 2019, pp. 647–657.
- [17] J. Bekker and J. Davis, "Learning from positive and unlabeled data: a survey," *Machine Learning*, pp. 719–760, 2020.
- [18] R. Kiryo, G. Niu, M. C. du Plessis, and M. Sugiyama, "Positive-unlabeled learning with non-negative risk estimator," in *Advances in neural information processing systems*, 2017, pp. 1675–1685.
- [19] T. Huang, S. You, F. Wang, C. Qian, C. Zhang, X. Wang, and C. Xu, "GreedyNASv2: Greedier search with a greedy path filter," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 902–11 911.
- [20] W. Hu, R. Le, B. Liu, F. Ji, J. Ma, D. Zhao, and R. Yan, "Predictive adversarial learning from positive and unlabeled data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, 2021, pp. 7806–7814.
- [21] S. Chang, B. Du, and L. Zhang, "Positive unlabeled learning with class-prior approximation," in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, 2021, pp. 2014–2021.
- [22] C. Gong, Q. Wang, T. Liu, B. Han, J. You, J. Yang, and D. Tao, "Instance-dependent positive and unlabeled learning with labeling bias estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4163–4177, 2021.
- [23] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," 2017.
- [24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
- [25] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015.
- [26] A. Radford, J. W. Kim, C. H. abd Aditya Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, 2021, pp. 8748–8763.
- [27] X. Sun, P. Hu, and K. Saenko, "Dualcoop: Fast adaptation to multi-label recognition with limited annotations," 2022.
- [28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.
- [29] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.
- [30] T. Huang, S. You, F. Wang, C. Qian, and C. Xu, "Knowledge distillation from a stronger teacher," in *Advances in Neural Information Processing Systems*, 2022, pp. 33 716–33 727.
- [31] T. Huang, S. You, B. Zhang, Y. Du, F. Wang, C. Qian, and C. Xu, "Dyrep: Bootstrapping training with dynamic re-parameterization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 588–597.
- [32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [33] S. Wen, W. Liu, Y. Yang, P. Zhou, Z. Guo, Z. Yan, Y. Chen, , and T. Huang, "Multilabel image classification via feature/label co-projection," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 11, pp. 7250–7259, 2020.