Exploring Structured Semantic Prior for Multi Label Recognition with Incomplete Labels

Supplementary Material

A. Methodology

Multi-label classification objective. We use the SPLC + Focal margin [9] loss to optimize the whole model. For an input image, we denote its visual representation as $\boldsymbol{f}$ and the label feature as $\boldsymbol{z_i}$, which can be derived by the image encoder and text encoder, respectively (see Eq. (5)). The multi-classification loss $\mathcal{L}_{cls}$ can be computed as

$$\begin{aligned}\mathcal{L}_{cls} = -\sum_{c=1}^{C}\Big\{ & y_c(1-p_c^m)^{\alpha}\log(p_c^m) + (1-y_c) \\ & \big[\mathbb{I}(p \le \beta){p_c}^{\alpha}\log(1-p_c) + \\ & (1-\mathbb{I}(p \le \beta))(1-p_c)^{\alpha}\log(p_c)\big]\Big\},\end{aligned} \tag{1}$$

where $p_c^m = \sigma(\text{sim}(\boldsymbol{f}, \boldsymbol{z_i})/\tau - m)$ is the likelihood and m is a margin parameter. α is set to 2 and β is a threshold to identify negative label.

B. Experiment Settings

B.1. Datesets for MLR with incomplete labels

For the single positive label setting, we conduct experiments on four standard benchmarks, *i.e.*, MS-COCO (COCO), PASCAL VOC 2012 (VOC), NUSWIDE (NUS) and CUB. The statistics of all benchmark datasets on training datasets are shown in Tab. 1. COCO contains 82,081 training images with 80 classes and a test set of 40,137 images. VOC consists of 5,717 training images with 20 classes and 5,823 images for test. NUS is a public multi-label image classification dataset which contains 269,648 images and each image is manually annotated with some of 81 categories. CUB consists of 5,994 training images covering 312 categories and 5,794 test images. For a fair comparison with [5], [9], we perform two different setups. The LargeLoss setup divides the training dataset into 80% for training and 20% for validation. The SPLC setup only trains on the overall training set and tests on the test set. The validation sets and test sets are always fully labeled.

For the partial label setting, we adopt three benchmarks, *i.e.*, MS-COCO (COCO), PASCAL VOC 2007 (VOC2007) and Visual Genome (VG-200). COCO dataset is same as the one used in the single positive label setting. VOC2007 contains a training set of 5,011 images and a test set of 4,952 images. VG-200 contains a total of 108,249 images covering tens of thousands of classes, most of which have only few samples. Following [7], we choose 200 frequent classes as the VG-200 subset, in which 10,000 images are randomly selected as the test set and the remaining 82,904 images are used as the training set.

B.2. Implementation details

For the single positive label setting, we use a single GPU with batch size 128. Each image of ours is uniformly resized to 224 × 224 while other methods resize to 448 × 448. We use the Adam optimizer and OneCycle learning rate schedule with the max learning rate of 3e-5. It is trained with 30 epochs in total.

For the partial label setting, we use two GPUs with batch size 32 and the max learning rate is 1e-5. We also train for 30 epochs on all benchmark datasets.

For data augmentation, we adopt the random horizontal flip and random resized crop for the weak transformation and the RandAugment for the strong transformation.

C. More experiments results

C.1. Model Analysis

Effect of different modules on the ResNet. To investigate the effectiveness of our proposed method with CNN-based MLR models, we conduct the ablation study on replacing the pretrained CLIP with the ResNet model. As shown in Tab. 2, each component can lead to performance improvement. Compared with the baseline, introducing SAM can achieve a performance improvement of 1.01%, which shows that exploiting the implicit label-to-label correspondence can benefit the MLR with incomplete labels. In addition, using the overall proposed PESSL can accomplish 0.53% mAP improvement, well demonstrating the advantage of incorporating the structured semantic prior to calibrate the semantic distribution. Finally, our proposed method significantly outperforms the baseline model with

Table 1. The statistics of all benchmark datasets on training sets.

Experiment setting	Dataset	Samples	Classes	Labels	Avg.label/img
Single positive label	COCO	82,081	80	241,035	2.9
	VOC	5,717	20	8,331	1.5
	NUS (LargeLoss)	150,000	81	284,611	1.9
	NUS (SPLC)	119,103	81	289,460	2.4
	CUB	5,994	312	188,343	31.4
Partial labels	COCO	82,081	80	241,035	2.9
	VOC2007	5,011	20	7,306	1.5
	VG-200	82,904	200	886,618	10.7
Real partial labels	OpenImages V3	3,552,103	5,000	13,440,371	3.8

Table 2. Analysis of different modules on the ResNet50 (%).

ResNet	SAM	PESSL	mAP
✓			73.18
✓	✓		74.19
✓	✓	✓	**74.72**

Table 3. Analysis on different combination of encoders (%). ResNet50 is pretrained on the ImageNet. ResNet50$_r$ is randomly initialized.

Image Encoder	Label Encoder	mAP
ResNet50	Linear	73.18
CLIP	Linear	73.33
ResNet50	CLIP	64.85
ResNet50$_r$	CLIP	41.71
CLIP	CLIP	**74.36**

1.54% improvement, well indicating the effectiveness and superiority of our method.

Combination of different encoders. We investigate different combinations of image encoders and label encoders. As shown in Tab. 3, compared with our method (the last row), replacing the CLIP label encoder with a linear prediction layer leads to inferior performance for a pretrained/CLIP-based ResNet50 (Row 1/2). Besides, replacing the CLIP image encoder with a pretrained/random ResNet50 (Row 3/4) also encounters great performance drops due to the destruction of image-label correspondence. This evidence shows the superiority of applying CLIP as the base MLR model.

Effect of the prompt learning. We investigate the effect of the prompt learning with the CMP model. As shown in Tab. 4, we can observe that with the random initialization, different prompt length can obtain similar performance. After leveraging the prompt template, *i.e.*, *a photo of a*, CMP with 4 prompts can achieve the best performance, indicating the positive effect of the hard prompt [8] as the prompt initialization.

Table 4. Effect of the prompt in the CMP (%).

Length	Initialization	mAP
4	template	**74.36**
4	Random	73.85
8	Random	73.87
16	Random	73.85

Table 5. Analysis on the semantic association module (SAM) (%).

Label Feature	$\boldsymbol{H}^L$	$\boldsymbol{H}^0, \boldsymbol{H}^L$	$\boldsymbol{H}^0 + \boldsymbol{H}^L$
	75.01	75.65	**76.42**
Correlation Matrix	Original	Sparse	Ours
	75.96	76.13	**76.42**

Analysis on SAM. For the proposed SAM component, we first investigate how to construct the label feature. We discuss three variants: 1) directly using $\boldsymbol{H}^L$, 2) using $\boldsymbol{H}^0$ and $\boldsymbol{H}^L$ in a multi-task learning strategy, and 3) the proposed residual connection, *i.e.*, $\boldsymbol{H}^0 + \boldsymbol{H}^L$. From Tab. 5, we can see that the residual connection obtains the best performance. We further analyze the effect of different label correlation matrix. We verify three strategies: 1) using original label correlation matrix (see Eq. (1)), 2) resulting in a sparse matrix by retaining the top K elements (see Eq. (2)), and 3) our adjusted the sparse matrix (see Eq. (4)), which is our proposed structured semantic prior, *i.e.*, $\boldsymbol{A}^*$. As shown in Tab. 5, our proposed method can achieve the state-of-the-art performance, demonstrating that the label-to-label correspondence can be well captured by the proposed structured semantic prior.

GT : bicycle, bus, car, person, traffic light.
Baseline : bicycle (0.95), car (0.97), person (0.81), traffic light (0.97), truck (0.62).
SCPNet : **bicycle (0.71)**, **bus (0.66)**, **car (0.85)**, **person (0.94)**, **traffic light (0.74)**.

GT : elephant, bird.
Baseline : elephant (0.97), sheep (0.83).
SCPNet : **elephant (0.94)**, **bird (0.78)**.

GT : bowl, oven, sink.
Baseline : oven (0.80), umbrella (0.80).
SCPNet : **bowl (0.57)**, **oven (0.80)**, **sink (0.64)**.

GT : motorcycle.
Baseline : none.
SCPNet : **motorcycle (0.97)**.

GT : broccoli, cake.
Baseline : none.
SCPNet : **broccoli (0.53)**, **cake (0.72)**.

GT : potted plant.
Baseline : none.
SCPNet : **potted plant (0.61)**.

Figure 1. Visualization of example results compared our SCPNet with the baseline model. Red color means results of false recognition and missing recognition. Blue color denotes ours results. GT means ground truth and the precision of recognition is in brackets.

Table 6. Different pseudo label selection of PESSL (%).

$\mathcal{L}_{cst}$	Soft Label	Threshold	Ours
mAP	75.88	75.80	**76.42**

Analysis of the pseudo label selection. We further investigate different pseudo label selection for weak transformation in the PESSL. We discuss three strategies: 1) using the weak transformation prediction probabilities as soft label, 2) setting threshold to filter the pseudo label, and 3) our method which further selects the top highest probability to construct a set of confident labels. As shown in Tab. 6, our method obtains the best performance, demonstrating that the selective construction of confidence labels is more compatible with the MLR task.

Stress-testing on domain-specific datasets. We conduct experiments on a common satellite dataset (AID[1]) and a common medical dataset (ChestX-ray14[2]) under the single positive label setting to perform stress-testing on domain-specific datasets which are far from those used in CLIP pretraining. As shown in Tab. 7, compared with CMP, our SCPNet can obtain 5.84% and 5.50% performance improvement on the AID and ChestX-ray14, respectively. These results show that our method achieves the best performance although CLIP cannot generalize well in these datasets, well demonstrating the generalization.

Table 7. Stress-testing on domain-specific datasets (%).

Method	AID	ChestX-ray14
SPLC [9]	71.26	25.60
CMP (ours)	67.48	22.42
SCPNet (ours)	**73.32**	**27.92**

DualCoOp vs. SCPNet. First, in terms of the model performance, for fair comparison, we implemented our method with a frozen image encoder on DualCoOp's code under the same setting, achieving 83.2% mAP (+1.3%). We also tuned DualCoOp under the SCPNet setting, where DualCoOp achieved inferior mAP with 82.4% (vs. ours: 83.8%). Second, in terms of the computation efficiency,

[1]https://github.com/Hua-YS/AID-Multilabel-Dataset
[2]https://nihcc.app.box.com/v/ChestXray-NIHCC

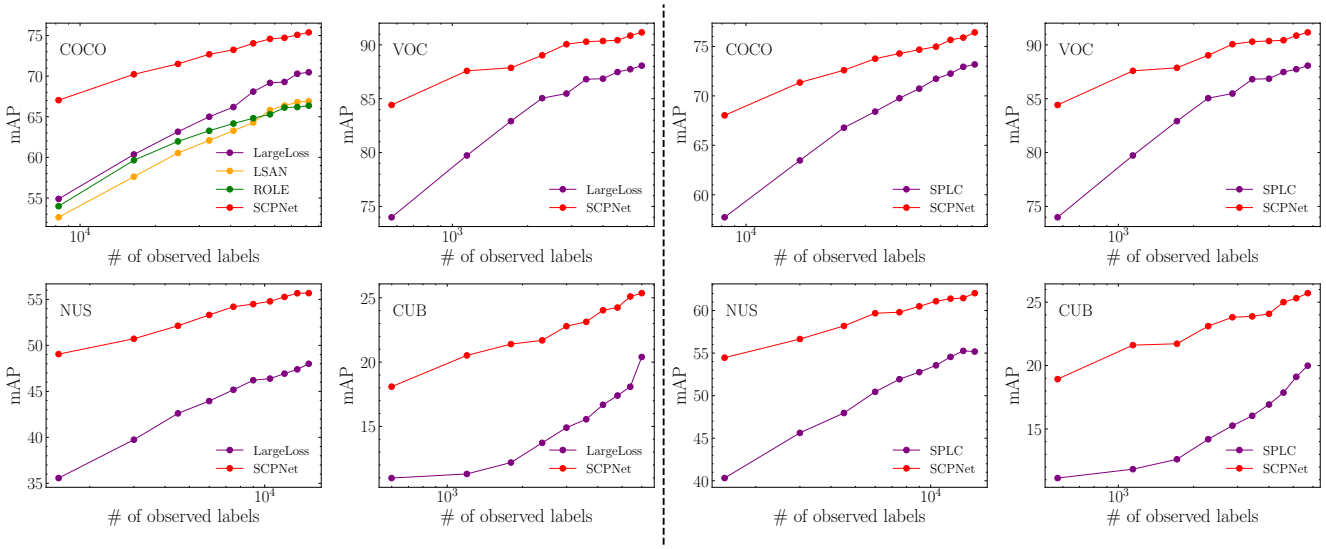

Figure 2. Results of the few-shot partial label setting on COCO, VOC, NUS and CUB dataset (left: in the LargeLoss setup, right: in the SPLC setup).

Table 8. DualCoOp vs. SCPNet in terms of computation cost

	DualCoOp	SCPNet
Training speed [iters/sec]	4.27	2.82
Trainable parameters	1.31M	3.41M
GPU memory for training	7.4G	9.8G
Inference speed [samples/sec]	318.59	322.76
GPU memory for inference	3.4G	3.4G
mAP performance on COCO	81.9%	83.2%

we compare our method with DualCoOp under the DualCoOp's setting. As shown in Tab. 8, during training, SCPNet consumes more resources than DualCoOp. But the required cost is not unaffordable in practice. During inference, both methods can derive label features offline. Therefore, SCPNet is comparable to DualCoOp in terms of computational cost while enjoying the superior performance. This shows SCPNet is more advanced or at least comparable when applied in practical scenarios.

These results clearly demonstrate the effectiveness and superiority of our method, compared to DualCoOp.

Hyper-parameters selection. For the single positive label setting, most hyper-parameters are directly borrowed from COCO, except for some dataset-dependent hyper-parameters, *e.g.*, K in Eq. (2) (best at 60/15/50/280 for COCO/VOC/NUS/CUB). s in Eq. (3) is empirically set to 0.2. For the partial label setting, most hyper-parameters are directly borrowed from the single positive label setting, except for the learning rate. Even so, our method can obtain consistent performance improvements in all scenarios, well demonstrating the robustness of hyper-parameters. We show that after more hyperparameter searches, we can obtain slightly better performance than the reported one, *e.g.*, 49.6% vs. 49.4% on VG-200.

C.2. Multi-label Recognition Results

Here we present the multi-label recognition results on the single positive label setting. As shown in Fig. 1, our proposed method can successfully recognize more accurate labels with lower false identifications (see examples in the first row). Besides, compared with the baseline model, our method can achieve fewer missing recognition for difficult labels, *e.g.*, "broccoli" in the middle of the second row. These results further demonstrate the effectiveness and the superiority of our proposed method.

C.3. Few-Shot Single Positive Label Setting

To investigate the effectiveness of the proposed method with a smaller number of training images, we further conduct the experiments in the few-shot single positive label setting under both the LargeLoss setup [5] and the SPLC setup [9].

As illustrated in Fig. 2 (left), in the LargeLoss setup, following [5], we randomly sample the training images from 10% to 100% and conduct experiments on the COCO dataset. We further compare our method with LargeLoss [5] on the other benchmark datasets, *i.e.*, VOC, NUS and CUB. Only given 10% of the training images, our method can obtain a maximal performance improvement of 12.16%, 7.47%, and 13.48% on COCO, VOC and NUS dataset, respectively. For CUB dataset, we achieve a maximal improvement of 9.22% with giving the training images of 20%. Overall, our method can accomplish an average

Table 9. Results with real partial label on OpenImage V3 dataset.

Method	G1	G2	G3	G4	G5	All Gs
CL [3]	70.4	71.3	76.2	80.5	86.8	77.1
IMCL [4]	71.0	72.6	77.6	81.8	87.3	78.1
Naive AN	77.1	78.7	81.5	84.1	88.1	82.0
WAN [2]	71.8	72.8	76.3	79.7	84.7	77.0
LSAN [2]	68.4	69.3	73.7	77.9	85.6	75.0
LargeLoss [5]	77.7	79.3	82.1	84.7	89.4	82.6
P-ASL [1]	73.2	78.6	85.1	87.7	90.6	83.0
SCPNet (ours)	**79.6**	**81.8**	**85.3**	**87.9**	**92.1**	**85.3**

performance improvement of 7.16%, 3.20%, 9.35%, and 7.62% on four datasets with the training images from 10% to 100%. Besides, as shown in Fig. 2 (right), in the SPLC setup, we present the comparison results with SPLC [9] on four benchmark datasets as well. The maximum performance improvement achieved by our method can reach 10.31%, 10.43% and 14.14% on COCO, VOC and NUS dateset, respectively. Our method can bring a maximum improvement of 9.78% with 20% training images on CUB dataset. Our SCPNet can obtain 5.06%, 4.79%, 8.76%, and 7.82% improvement on average for the four datasets, respectively.

These experimental results show that our proposed SCPNet can significantly achieve state-of-the-art performance in different few-shot single positive label setting, well indicating the generalization and superiority.

C.4. Real Partial Label Scenario

Dataset and implementation details. To analyze the effectiveness of the proposed method in real partial label scenario, we conduct experiments on the OpenImage V3 [6] dataset with 5,000 classes. The details are shown in Tab. 1, OpenImage V3 contains 3.5M training images, 42k validation images, and 125k test images. Follwing [5], we divide the training images into 5 groups, where G1 has the smallest number of the counted images and G5 is the largest one. All Gs corresponds to the set of all categories.

Compared methods. We compare our method with Curriculumn Labeling(CL) [3], IMCL [4], Naive AN, Weak AN (WAN) [2], Label Smoothing with AN (LSAN) [2], LargeLoss [5] and P-ASL [1].

Results. As shown in Tab. 9, our method outperforms the state-of-the-art methods on G1, G2, G3, G4 and G5 with an improvement of 1.9%, 2.5%, 0.2%, 0.2% and 1.5%, respectively. As a whole, our proposed SCPNet can accomplish a performance improvement of 2.3% on all Gs, well demonstrating the effectiveness and generalization to practical scenarios.

References

[1] 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. 5

[2] Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, and Nebojsa Jojic. Multi-label learning from single positive labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 933–942, 2021. 5

[3] 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. 5

[4] Dat Huynh and Ehsan Elhamifar. Interactive multi-label cnn learning with partial labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9423–9432, 2020. 5

[5] Youngwook Kim, Jae Myung Kim, Zeynep Akata, and Jungwoo Lee. Large loss matters in weakly supervised multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14156–14165, 2022. 1, 4, 5

[6] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2(3):18, 2017. 5

[7] Tao Pu, Tianshui Chen, Hefeng Wu, and Liang Lin. Semantic-aware representation blending for multi-label image recognition with partial labels. *arXiv preprint arXiv:2203.02172*, 2022. 1

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[9] 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. 1, 3, 4, 5