Hierarchical Prompt Learning Using CLIP for Multi-label Classification with Single Positive Labels

Ao Wang
wa22@mails.tsinghua.edu.cn
Tsinghua University
School of Software, BNRist
Beijing, China

Hui Chen*
jichenhui2012@gmail.com
Tsinghua University
BNRist
Beijing, China

Zijia Lin
linzijia07@tsinghua.org.cn
Tsinghua University
Beijing, China

Zixuan Ding
dingzixuan96@163.com
Xidian University & Zhuoxi Institute of Brain and Intelligence
Xi'an & Hangzhou, China

Pengzhang Liu
liupengzhang@jd.com
jd.com
Beijing, China

Yongjun Bao
baoyongjun@jd.com
jd.com
Beijing, China

Weipeng Yan
paul.yan@jd.com
jd.com
Beijing, China

Guiguang Ding*
dinggg@tsinghua.edu.cn
Tsinghua University
School of Software, BNRist
Beijing, China

ABSTRACT

Collecting full annotations to construct multi-label datasets is difficult and labor-consuming. As an effective solution to relieve the annotation burden, single positive multi-label learning (SPML) draws increasing attention from both academia and industry. It only annotates each image with one positive label, leaving other labels unobserved. Therefore, existing methods strive to explore the cue of unobserved labels to compensate for the insufficiency of label supervision. Though achieving promising performance, they generally consider labels independently, leaving out the inherent hierarchical semantic relationship among labels which reveals that labels can be clustered into groups. In this paper, we propose a hierarchical prompt learning method with a novel Hierarchical Semantic Prompt Network (HSPNet) to harness such hierarchical semantic relationships using a large-scale pretrained vision and language model, i.e., CLIP, for SPML. We first introduce a Hierarchical Conditional Prompt (HCP) strategy to grasp the hierarchical *label-group* dependency. Then we equip a Hierarchical Graph Convolutional Network (HGCN) to capture the high-order *inter-label* and *inter-group* dependencies. Comprehensive experiments and analyses on several benchmark datasets show that our method significantly outperforms the state-of-the-art methods, well demonstrating its superiority and effectiveness. Our code will be available at https://github.com/jameslahm/HSPNet.

*Corresponding author.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3611988

CCS CONCEPTS

• **Computing methodologies** → **Object recognition**.

KEYWORDS

multi-label classification, image recognition, vision and language, weak supervision, hierarchical relationship

ACM Reference Format:
Ao Wang, Hui Chen, Zijia Lin, Zixuan Ding, Pengzhang Liu, Yongjun Bao, Weipeng Yan, and Guiguang Ding. 2023. Hierarchical Prompt Learning Using CLIP for Multi-label Classification with Single Positive Labels. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3581783.3611988

1 INTRODUCTION

Multi-label classification (MLC) aims to describe the image content with multiple semantic labels. Benefiting from the collected multi-label datasets with full annotations, MLC has achieved remarkable progress in recent years [18, 24, 34, 40]. However, when the label set scales up, collecting full annotations for large-scale datasets becomes very difficult and labor-consuming. Therefore, a full labeling strategy encounters a huge limitation in the practical application of MLC. To mitigate this issue, researchers have dedicated themselves to a new setting, i.e., single positive multi-label learning (SPML), in which merely one positive label is annotated for each image [5, 11, 12, 35]. Such an extreme setting has shown great superiority in reducing the annotation burden [36], thus drawing increasing attention from both academia and industry. However, compared with the full label setting, the insufficiency of supervision causes great challenges to SPML [5].

Existing works strive to tackle such an insufficiency issue by mining knowledge from the unobserved labels, aiming to recover more unlabelled but potentially true labels. Nonetheless, most works

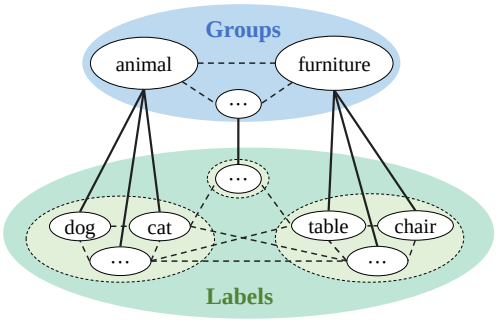

Figure 1: The hierarchical semantic relationships among labels. Solid lines denote the explicit *label-group* dependency in the hierarchical semantic structure. Dashed lines are the implicit *inter-label* and *inter-group* dependencies.

simply explore the cue of unannotated labels independently, without consideration of the semantic relationships among labels. For example, [13, 41] propose to correct the loss for labels individually according to the model's predicted probability. [42] present an entropy-maximization loss to acknowledge each unannotated label separately.

It's a consensus that there generally exist rich semantic relationships among different labels [3, 7, 45]. For example, "table" and "chair" are likely to appear together. "fish" is not likely to co-occur with "sky". Intuitively, unobserved labels can be reasoned from observed labels by leveraging their semantic relationships, thus facilitating mining more supervision information. Such label dependencies have been explored under the full label setting [3] and the partial label setting [7] with graph neural networks [33]. However, rare works pay attention to leveraging such dependencies to enhance the SPML performance. The main reason is that the severe deficiency of label annotations in SPML makes it hard to establish sufficient label dependencies to discover more unobserved true labels.

Recently, large-scale vision-language pretrained models such as the Contrastive Language-Image Pre-training (CLIP) [23], have become one of hottest topics in both academic and industrial communities. By exploring contrastive learning with about 400 million noisy image-text pairs, CLIP can well exploit the semantic relationships between images and their associated texts. Thanks to the language supervision, models based on CLIP achieve impressive results on various downstream vision tasks [19, 27, 37]. For example, [44] propose a context optimization method, dubbed CoOp, to adapt CLIP to the single-label classification, achieving impressive classification performance. [27] present dual prompts, i.e., a positive prompt and a negative one, to transfer the knowledge in CLIP to multi-label classification. Despite great performance, existing methods, e.g., [27, 44], simply strive to leverage the label semantics in CLIP, taking no consideration of the abundant semantic dependencies, which is convincingly valuable in SPML.

To harness the promising label dependency in CLIP for SPML, we are motivated by an intrinsic characteristic, i.e., the hierarchical semantic relationship, which indicates that labels can be clustered into groups, as shown in Figure 1. We can observe: 1) the *label–group* dependency: each label can be associated with one group according to a certain measurement, e.g., the semantic similarity. For example, "cat" and "dog" belong to the "animal" group. 2) the *inter-label* dependency: semantically related labels, which can be within or across groups, can co-occur in the same image; 3) and the *inter-group* dependency: among groups, there also implicitly exist dependencies that may indicate the co-occurrence. Intuitively, labels in the same group are more likely to be reasoned from the given positive label than those in other groups that are less semantically related to the positive label. Therefore, leveraging the label-group correspondence, models can conveniently grasp related unobserved labels within an identical group and pay more attention to less related inter-group labels, thus relieving the burden of mining complicated label dependencies. Therefore, **compared with simple label dependencies, such hierarchical semantic relationships can provide richer dependencies, thus being notably favorable for SPML**.

In this paper, we propose a hierarchical prompt learning method by a **H**ierarchical **S**emantic **P**rompt Network, namely HSPNet, which aims to explore the valuable hierarchical semantic relationship using CLIP for SPML. Specifically, we first introduce a **H**ierarchical **C**onditional **P**rompt (HCP) strategy to efficiently adapt CLIP to the downstream task of SPML. By explicitly conditioning label prompts on their associated group representation, the proposed HCP can well grasp the *label-group* dependency, ending up with compact group-aware label features. Besides, we propose a **H**ierarchical **G**raph **C**onvolutional **N**etwork (HGCN) to comprehensively capture the high-order *inter-label* and *inter-group* relationships. Thanks to HGCN, our method can enable subtle semantic associations, thus enhancing the label hypothesis prediction.

To verify the effectiveness of our proposed HSPNet, we conduct extensive experiments on a series of widely used benchmark datasets for SPML, i.e. MS COCO [17], NUSWIDE [4], CUB [29], and Pascal VOC [8]. Experimental results show that our method can significantly outperform the state-of-the-art methods on all datasets, well demonstrating its effectiveness and superiority. We further reveal the remarkable generalization capability of HSPNet in other practical MLC scenarios, e.g., the few-shot SPML setting and the partial label setting, where our method also obtains consistent performance gains compared to other methods. Additionally, we investigate the efficacy of HSPNet in scenarios with a lack of explicit group information, by clustering label features to automatically construct hierarchical relationships. The experimental results well demonstrate its practicality in such scenarios.

To sum up, our contributions are three folds:

- We propose a Hierarchical Semantic Prompt Network (HSPNet) which can effectively explore the hierarchical semantic relationship in the vision-language pretrained model, i.e., CLIP, to compensate for the dilemma of poor supervision in SPML.
- We design a Hierarchical Conditional Prompt (HCP) and a Hierarchical Graph Convolutional Network (HGCN), which can efficiently incorporate the label-group dependency in the hierarchical prompt learning and capture the high-order inter-label and inter-group relationships, respectively.

- We conduct extensive experiments and analyses on widely used benchmark datasets, which show our method can consistently achieve state-of-the-art performance, well demonstrating its effectiveness and superiority.

2 RELATED WORK

Multi-label classification with single positive labels. Thanks to the rapid development of deep learning and the emergence of large-scale multi-label datasets, multi-label classification under different settings has achieved great progress in past years. Various improvements for loss function [24] and backbone networks [18] obtain notable success under the full label setting and the partial label setting. However, they are not perfectly applicable in SPML due to the special limitation of only one positive label for each image. To mitigate the issue of severely weak label supervision, different robust losses are designed for SPML. [41] present a Hill loss and a self-paced loss correction method to alleviate the effect of false negative labels. To relieve the memorization effect during training in SPML, [13] further propose to correct or reject the large loss samples. [42] then design an entropy-maximization loss to attain a special gradient regime for unobserved labels. In addition to the robust loss design, different training schemes are also proposed to increase the pseudo-label annotations. [28] design a scheme to ignore the labels expected to be unobserved positives. [42] further propose an asymmetric pseudo-labeling strategy to generate more precise supervision. Though previous works in SPML achieve notable performance gains, the label dependencies are generally ignored, resulting in limitations in alleviating the deficiency of label supervision.

Mining label dependencies for MLC. Considering associated objects normally co-occur in one image, capturing label dependencies to discover more positive labels is actively explored in the full label setting [18, 31] and the partial label setting [7, 45]. However, adapting existing methods to SPML is limited mainly due to the few annotations. For example, [3, 31] adopt the conditional probability matrix to model the label dependency explicitly, while such statistic information depends on adequate annotations. [30, 38] present the CNN-RNN architecture and [18] leverage the transformer decoder structure to capture the label co-occurrence dependency implicitly, which, however, require sufficient label supervision for training. Moreover, the *hierarchical* semantic relationships which provide richer label dependencies are rarely explored in MLC. Therefore, different from previous works, we aim to introduce such hierarchical relationships into SPML without the dependence on sufficient label annotations, which is notably valuable in SPML.

Vision-language models in downstream visual tasks. Recently, vision-language pretrained models have obtained remarkable success in various downstream visual tasks [15, 23, 32, 39]. Therefore, researchers actively explore how to leverage the abundant prior knowledge in vision-language pretrained models [9, 10, 19, 26, 27, 37]. [37] propose a Dual-Modal decoder to align visual and textual features for zero-shot multi-label classification. Additionally, [19] present fusioner to pair the visual representation and language concept for adapting the vision-language pretrained models to the task of open-vocabulary semantic segmentation. Dual prompts are used in [27] for MLC. They present a positive prompt and a negative one to transfer the knowledge in CLIP. However, they discard the hierarchical semantic relationship among labels and thus not optimal for exploring CLIP's knowledge.

3 METHODOLOGY

In this section, we describe the proposed hierarchical semantic prompt network (HSPNet) in detail. Our HSPNet aims to explore the hierarchical semantic relationship using CLIP for SPML. To this end, we first extract cross-modality features for the input image and labels using CLIP. Then, we propose a hierarchical conditional prompt strategy to adapt CLIP to SPML efficiently via incorporating the hierarchical label-group structure (see Section 3.1). To further enhance the semantic representation with high-order hierarchical semantic dependencies, we adopt a hierarchical graph convolutional network to refine the label features and the group features with inter-label and inter-group relationships (see Section 3.2). Figure 2 illustrates the overview of our HSPNet.

3.1 Hierarchical Conditional Prompt

Recently, the development of large-scale cross-modality pretrained models draws great attention from the vision community. Researchers investigate different prompt learning methods to efficiently adapt such powerful large-scale models to downstream vision tasks, e.g., CoOp [44], CoCoOp [43]. However, such a simple prompt strategy discards the rich hierarchical semantic dependency among labels, thus being sub-optimal for SPML. Here, we present a hierarchical conditional prompt (HCP) strategy to explore the hierarchical dependency between groups and labels. We encourage the SPML model to learn prompts for groups and labels simultaneously. Then, during the construction of prompts for labels, we optimize the label prompt conditioned upon its corresponding group with the guidance of the label-group dependency.

Specifically, given a label set $Y = \{y_0, y_1, ..., y_n\}$, we first obtain their hierarchical structure like in Figure 1 by the provided concept taxonomy in datasets or by manually dividing the labels into groups based on their feature similarity or the semantic relationship in the WordNet [20]. Then, we derive the groups $Z = \{z_1, z_2, ..., z_m\}$ and the surjective function $\Phi : Y \rightarrow Z$ by the hierarchical structure. Through Φ, we can obtain a label set h_j in which all labels belong to the group z_j by:

$$h_j = \{y \in Y | \Phi(y) = z_j\} \tag{1}$$

Considering a CLIP model with a text encoder, denoted as $E(\cdot)$ and an image encoder, denoted as $F(\cdot)$, we introduce a fixed prompt template, i.e., *a photo of a [CLS]* where *[CLS]* is the name of a given group, to extract the group features. For ease of explanation, we denote the word embeddings of the prefix "*a photo of a*" as $\boldsymbol{o}_j$, which is generated by the CLIP model. Then, for each group z_j, the group feature $\boldsymbol{g}_j$ of z_j can be derived via the text encoder of CLIP by:

$$\boldsymbol{g}_j = E([\boldsymbol{o}_j, \boldsymbol{d}_j]) \tag{2}$$

where $\boldsymbol{d}_j$ is the word embedding of z_j in the pretrained CLIP.

To incorporate the label-group dependency into the extraction of label features, we dedicatedly leverage the hierarchical semantic structure in the construction of label prompts. We design a group-aware meta network to generate the label prompts based on $\boldsymbol{G}$ =

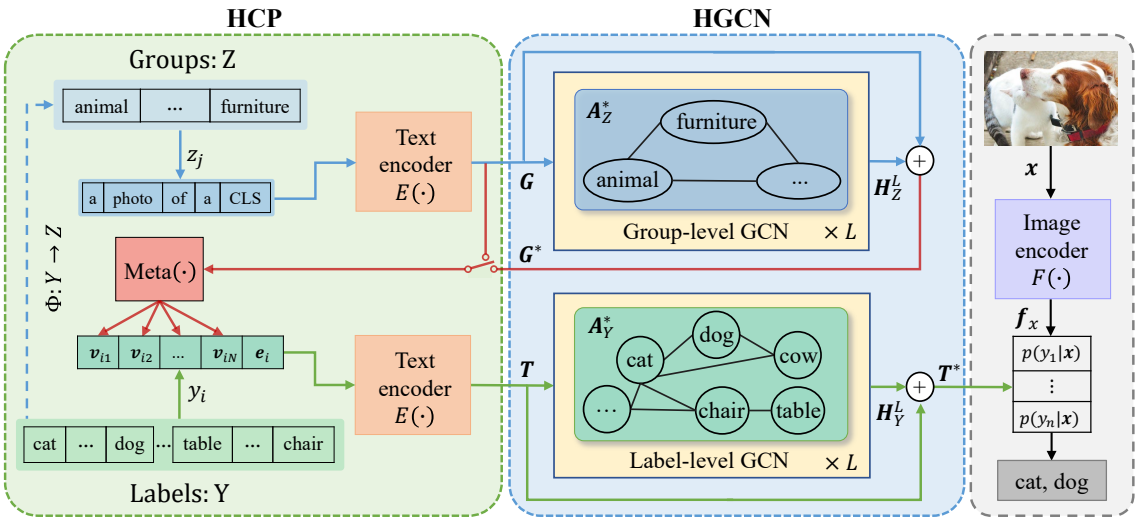

Figure 2: The overview of the proposed HSPNet. We design a hierarchical conditional prompt (HCP) strategy to explore the inherent *label-group* dependency between labels and groups. A hierarchical graph convolutional network (HGCN) is introduced to grasp the *inter-label* and *inter-group* dependencies.

$\{\boldsymbol{g}_1, \boldsymbol{g}_2, ..., \boldsymbol{g}_m\}$, where $\boldsymbol{G}$ is a combination of group features derived from Equation (2). The group-aware meta network is a lightweight neural network, which is composed of two linear layers and one activation layer, i.e. Linear-ReLU-Linear. We denote the network as Meta$(\cdot)$. For each label $y_i \in h_j$ where h_j is a label set corresponding to the group z_j (see Equation (1)), we can obtain its group-aware label prompt with N learnable soft prompt tokens conditioned on its corresponding group feature $\boldsymbol{g}_j$ by:

$$\boldsymbol{s}_i = \{\boldsymbol{v}_{i1}, \boldsymbol{v}_{i2}, ..., \boldsymbol{v}_{iN}\} = \text{Meta}(\boldsymbol{g}_j) \tag{3}$$

where $\boldsymbol{s}_i$ denotes the group-aware label prompt of y_i. Then, we can derive the group-aware label feature $\boldsymbol{t}_i$ of y_i by:

$$\boldsymbol{t}_i = E([\boldsymbol{s}_i, \boldsymbol{e}_i]) \tag{4}$$

where $\boldsymbol{e}_i$ is the word embedding of y_i in the pretrained CLIP.

As illustrated by Equation (3), the label prompt for label y_i, i.e., $\boldsymbol{s}_i$, is conditioned on the group representation, i.e., $\boldsymbol{g}_j$. For all $y_i \in h_j$ corresponding to group z_j, they share the same group feature, which naturally encourages label features belonging to the same group to be clustered in the same manifold space, leading to more compact group-aware feature clusters for labels. Such a property can enable SPML models to conveniently associate semantically related labels with the input image, relieving the burden of image-label matching during training.

3.2 Hierarchical Graph Convolutional Network

As the HCP mainly leverages label-group relationships, the hierarchical inter-label and inter-group relationships are not adequately explored. Therefore, we present a hierarchical graph convolutional network (HGCN) to further comprehensively capture the high-order inter-label and inter-group relationships. We first capture the inter-label and inter-group relationships in the form of dependency graphs. Then, two GCN modules are employed to refine the group features, i.e., $\boldsymbol{G} = \{\boldsymbol{g}_1, \boldsymbol{g}_2, ..., \boldsymbol{g}_m\}$ derived by Equation (2), and the label features, i.e., $\boldsymbol{T} = \{\boldsymbol{t}_1, \boldsymbol{t}_2, ..., \boldsymbol{t}_n\}$ derived by Equation (4), with the guidance of the group dependency graph and the label dependency graph, respectively.

Given groups $Z = \{z_1, z_2, ..., z_m\}$ and their features $\boldsymbol{G} = \{\boldsymbol{g}_1, \boldsymbol{g}_2, ..., \boldsymbol{g}_m\}$, we can obtain the group correlation matrix $\boldsymbol{A}_Z = (a_{ij})_{m\times m}$ by:

$$a_{ij} = \text{sim}(\boldsymbol{g}_i, \boldsymbol{g}_j) \tag{5}$$

where sim$(\cdot,\cdot)$ denotes the cosine similarity. The correlation matrix $\boldsymbol{A}_Z$ forms a graph, where nodes represent the groups Z and each edge with the weight a_{ij} indicates the semantic relationship between group z_i and group z_j.

To eliminate the noise inside the correlation prior, the top K elements are reserved and the remaining elements are set to zero for each row $\boldsymbol{a}_i$ in $\boldsymbol{A}_Z$, resulting in a sparse matrix $\overline{\boldsymbol{A}}_Z = (\overline{a}_{ij})_{m\times m}$ formulated as:

$$\overline{a}_{ij} = \begin{cases} a_{ij}, & if\ j \in \text{topK}(\boldsymbol{a}_i) \\ 0, & if\ j \notin \text{topK}(\boldsymbol{a}_i) \end{cases} \tag{6}$$

To alleviate the over-smoothing problem, we derive the final group dependency graph $\boldsymbol{A}_Z^* = (a_{ij}^*)_{m\times m}$ similar to [3] by:

$$a_{ij}^* = \begin{cases} \overline{a}_{ij} \cdot r / \sum_{j' \neq i}^{m} \overline{a}_{ij'}, & if\ i \neq j \\ 1 - r, & if\ i = j \end{cases} \tag{7}$$

where r is the hyper-parameter that determines the sum of weights assigned to neighboring groups. $\boldsymbol{A}_Z^*$ encodes the structural correspondence of groups via the weights representing the semantic relationships, which then can be used to refine the group features. Given the label set $Y = \{y_1, y_2, ..., y_n\}$, the label dependency graph $\boldsymbol{A}_Y^*$ can be derived following the same procedure as $\boldsymbol{A}_Z^*$.

With the guidance of the hierarchical dependency graph, i.e., $\boldsymbol{A}_Z^*$ and $\boldsymbol{A}_Y^*$, we employ the graph convolutional network (GCN) with L

layers to progressively refine the group features $\boldsymbol{G} = \{\boldsymbol{g}_1, \boldsymbol{g}_2, ..., \boldsymbol{g}_m\}$ and the label features $\boldsymbol{T} = \{\boldsymbol{t}_1, \boldsymbol{t}_2, ..., \boldsymbol{t}_n\}$, which are denoted as the group-level GCN and the label-level GCN, respectively. For groups, taking $\boldsymbol{H}_Z^0 = \boldsymbol{G}$ as the input features, the l-th GCN layer is updated as follows:

$$\boldsymbol{H}_Z^{l+1} = \rho(\boldsymbol{A}_Z^* \boldsymbol{H}_Z^l \boldsymbol{W}_Z^l) \tag{8}$$

where ρ is the activation function and $\boldsymbol{W}_Z^l$ is the learnable parameters in the l-th layer. Then, we can obtain the refined group features via a residual connection, i.e., $\boldsymbol{G}^* = \boldsymbol{H}_Z^0 + \boldsymbol{H}_Z^L$. For labels, given the label features $\boldsymbol{T}$ and label dependency graph $\boldsymbol{A}_Y^*$, we also utilize L GCN layers to refine the input features $\boldsymbol{H}_Y^0 = \boldsymbol{T}$, in which the l-th GCN layer is updated by:

$$\boldsymbol{H}_Y^{l+1} = \rho(\boldsymbol{A}_Y^* \boldsymbol{H}_Y^l \boldsymbol{W}_Y^l) \tag{9}$$

where ρ is the activation function and $\boldsymbol{W}_Y^l$ is the l-th layer's learnable parameters. We obtain the refined label features $\boldsymbol{T}^* = \{\boldsymbol{t}_1^*, \boldsymbol{t}_2^*, ..., \boldsymbol{t}_n^*\}$ by $\boldsymbol{T}^* = \boldsymbol{H}_Y^0 + \boldsymbol{H}_Y^L$.

Thanks to the GCN, the group and label features can be progressively refined through the hierarchical semantic relationships represented by $\boldsymbol{A}_Z^*$ and $\boldsymbol{A}_Y^*$. Thus, semantically similar labels and groups will be closer in the feature space, while semantically different labels and groups will obtain more discriminative features. Refined group features can generate better label prompts and refined label features can further lead to stronger matching between the image and its related labels, which is desired in SPML.

3.3 Training Objective

For an image $\boldsymbol{x}$, we obtain its feature $\boldsymbol{f}_x$ via the image encoder $F(\cdot)$ in CLIP as $\boldsymbol{f}_x = F(\boldsymbol{x})$. The predicted probability $p(y_i|\boldsymbol{x})$ that the image $\boldsymbol{x}$ contains the label y_i can be computed as:

$$p(y_i|\boldsymbol{x}) = \sigma(\text{sim}(\boldsymbol{f}_x, \boldsymbol{f}_{y_i})/\tau) \tag{10}$$

where $\boldsymbol{f}_{y_i}$ can be the label feature $\boldsymbol{t}_i$ of y_i obtained by Equation (4) or the refined label feature $\boldsymbol{t}_i^*$ derived by Equation (9). σ is the sigmoid function. τ is the learnable temperature parameter in CLIP. $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity.

During training, we adopt the same objective as the one in [41] to optimize the HSPNet:

$$\begin{aligned} L = -\sum_{i=1}^{n} \{ & y_i(1-p_i(m'))^{\alpha} log(p_i(m')) + (1-y_i) \\ & [\mathbb{I}(p_i \le \beta) p_i^{\alpha} log(1-p_i) + \\ & (1-\mathbb{I}(p_i \le \beta))(1-p_i)^{\alpha} log(p_i)] \} \end{aligned}$$

where $p_i(m') = \sigma(\text{sim}(\boldsymbol{f}_x, \boldsymbol{f}_{y_i})/\tau - m')$ and p_i denotes the predicted probability $p(y_i|\boldsymbol{x})$. α is empirically set to 2. m' is the margin parameter which is set to 1 by default. β is a threshold used to correct annotations for unobserved labels, which is empirically set to 0.6.

Since the proposed HCP and HGCN are separate from the image encoder, we can conveniently derive label features, i.e., $\boldsymbol{T}$ or $\boldsymbol{T}^*$, as decision boundaries for SPML after training. Therefore, during inference, the HSPNet is equivalent to a convolutional neural network, e.g., ResNet50. As a result, no extra computational cost is brought by the proposed HSPNet.

4 EXPERIMENT

4.1 Experimental Setup

Datasets. For SPML, we conduct experiments on benchmark datasets, i.e., COCO [17], PASCAL VOC 2012 (VOC) [8], NUSWIDE (NUS) [4], and CUB [29]. Considering that one category and multiple attributes are annotated for each image in the CUB dataset, we classify the images according to the attributes rather than bird categories for CUB. Following [13, 41], we reserve one label in the training set for multi-label datasets with full annotations to simulate the single positive label setting. To fairly compare with the state-of-the-art methods, we perform two different setups, i.e., the LargeLoss setup and SPLC setup, according to [13] and [41], respectively. In the LargeLoss setup, we divide the training set into 80% for training and 20% for validation. For the SPLC setup, we train on the full training set for all datasets.

Implementation details. For fair comparisons with other methods, we adopt the ResNet50-based CLIP initialized by the published pretrained weights. We use a single GPU with a batch size of 128. Following [13, 41], we employ the widely used random horizontal flip and random resized crop for data augmentation. Besides, we adopt the Adam optimizer and OneCycle learning rate schedule with $1e^{-4}$ as the maximal learning rate during training. The mean average precision (mAP) is adopted as the evaluation metric by default. We directly use the label-group correspondence provided in the original datasets. Details of the hierarchical structure for different benchmark datasets are provided in Appendix A due to the space limit.

4.2 Comparisons with State-of-the-Arts for SPML

We compare our method with existing state-of-the-art methods for SPML, including LSAN [5], ROLE [5], LargeLoss [13], Hill [41], and SPLC [41]. To fairly verify the effectiveness of our method, we employ their published codes of the state-of-the-art methods, i.e., LargeLoss [13] and SPLC [41], and re-implement them with the same ResNet50-based CLIP as ours but with the single prompt learning strategy like CoOp [44]. Their results are denoted as LargeLoss* and SPLC*, respectively. We also apply the state-of-the-art model for the partial label setting, i.e., DualCoop [27], in the SPML using their published code.

As shown in Table 1, for both setups, our method can significantly outperform existing state-of-the-art methods on all benchmark datasets. Specifically, in the LargeLoss setup, compared with the second best result, the proposed method can achieve a maximal performance improvement of 1.4% (COCO). In the SPLC setup, the maximal performance improvement obtained by our method can reach 1.6% (CUB). As a whole, our method can accomplish an overall improvement of 1.9% in both setups, which indicates the effectiveness of our method. Note that although DualCoop achieves the state-of-the-art performance for the partial label setting (seeing Table 7), it does not adapt well in the difficult SPML setting, which we attribute to insufficient positive and negative supervision in SPML. In contrast, the proposed HSPNet can effectively explore the hierarchical semantic relationship to enhance the label correspondence, making it a promising method for SPML.

Table 1: Comparison results in SPML (%). Results in bold are the best performance and those with the underline are the second best. * indicates models using CLIP weights as ours.

Method	LargeLoss setup					SPLC setup				
	COCO	VOC	NUS	CUB	Avg.	COCO	VOC	NUS	CUB	Avg.
LSAN [5]	69.2	86.7	50.5	17.9	56.1	70.5	87.2	52.5	18.9	57.3
LargeLoss [13]	71.6	<u>89.3</u>	49.6	21.8	58.1	-	-	-	-	-
Hill [41]	-	-	-	-	-	73.2	87.8	55.0	18.8	58.7
SPLC [41]	72.0	87.7	49.8	18.0	56.9	73.2	88.1	55.2	20.0	59.1
LargeLoss*	72.9	88.1	52.9	<u>22.4</u>	<u>59.1</u>	74.0	89.3	58.5	<u>22.7</u>	61.1
SPLC*	<u>73.4</u>	87.4	<u>55.1</u>	20.1	59.0	<u>74.4</u>	88.5	<u>60.7</u>	21.4	<u>61.2</u>
DualCoop*	72.9	87.8	50.6	21.0	58.1	73.5	<u>89.5</u>	55.9	21.9	60.2
HSPNet (ours)	**74.8**	**89.4**	**56.3**	**23.4**	**61.0**	**75.7**	**90.4**	**61.8**	**24.3**	**63.1**

Table 2: Quantitative results for the effect of different modules in our proposed HSPNet for both setups (%).

Model	LargeLoss Setup					SPLC Setup				
	COCO	VOC	NUS	CUB	Avg.	COCO	VOC	NUS	CUB	Avg.
SPL	73.39	87.35	55.09	20.12	58.99	74.36	88.46	60.66	21.42	61.22
SPL+GCN	74.14	87.93	55.49	20.34	59.48	75.12	89.09	61.08	21.66	61.74
HCP	74.36	88.91	55.79	21.69	60.19	75.18	89.84	61.20	23.65	62.47
HCP+HGCN (i.e., HSPNet)	**74.83**	**89.44**	**56.33**	**23.43**	**61.01**	**75.71**	**90.40**	**61.75**	**24.34**	**63.05**

4.3 Ablation Study

In order to analyze the effectiveness of each module, we conduct the ablation study in both the LargeLoss setup and the SPLC setup. To verify the superiority of exploring the hierarchical semantic relationship, we introduce two baseline methods: 1) a single prompt learning (SPL) which directly tunes the learnable prompt tokens like CoOp [44]; 2) SPL+GCN in which we use L GCN layers to refine the label features like Equation (9) on top of SPL. Only the inter-label relationship is captured in SPL+GCN. For both baseline methods, we initialize the SPML model with the same CLIP weights as those used in our HCP and HSPNet for fair comparisons.

As shown in Table 2, each module in the proposed method can consistently achieve performance improvements on all datasets with different setups. Specifically, benefiting from incorporating the label-group dependency in prompt learning, the designed HCP can obtain overall improvements of 1.2% and 1.25% mAP, compared with SPL in the LargeLoss setup and SPLC setup, respectively. These results demonstrate the superiority of adapting the vision-language model, i.e., CLIP, to SPML by the proposed group-aware label prompt. On top of HCP, HCP+HGCN can further lead to 0.82% and 0.58% mAP gains in the LargeLoss setup and SPLC setup, respectively. Such improvements can be attributed to simultaneously refining the label and group features with the hierarchical semantic relationship in the introduced HGCN. Furthermore, the proposed HSPNet can significantly outperform the SPL+GCN model by 1.53% and 1.31% in terms of mAP in both setups, respectively. It well demonstrates the strength of exploring the hierarchical semantic relationship among labels in the task of multi-label classification with single positive labels.

Table 3: Analyses on the label-group dependency (%).

SPL+GCN	w/o LGP	w/o GRP	Ours
75.12	75.38	75.41	**75.71**

4.4 Model Analyses

In this section, we perform comprehensive analyses for our proposed method. We conduct the experiments on the COCO dataset in the SPLC setup following [41].

Analyses on the label-group dependency. The label-group dependency determines how to assign labels to their corresponding groups. Here, we inspect the effects of two circumstances: (1) labels and groups are misaligned, and (2) semantic relationships among groups are not well captured. Therefore, we randomly assign n labels to m groups without the label-group structure prior provided by the benchmark dataset, denoted as "w/o LGP (Label Group Prior)". We change the group names to "group1", "group2", ..., "group$m$" to eliminate the semantic relationship prior among groups in the pretrained CLIP model, denoted as "w/o GRP (Group Relationship Prior)". We employ the HSPNet for inspections. As shown in Table 3, compared with "w/o LGP" and "w/o GRP", our method can consistently obtain better performance, indicating that obtaining accurate label-group dependency is beneficial.

Analyses on automated hierarchical dependencies. In practical scenarios, explicit group information can be hard to obtain especially when there are a large number of labels. Under this circumstance, we propose to cluster label features and use the derived

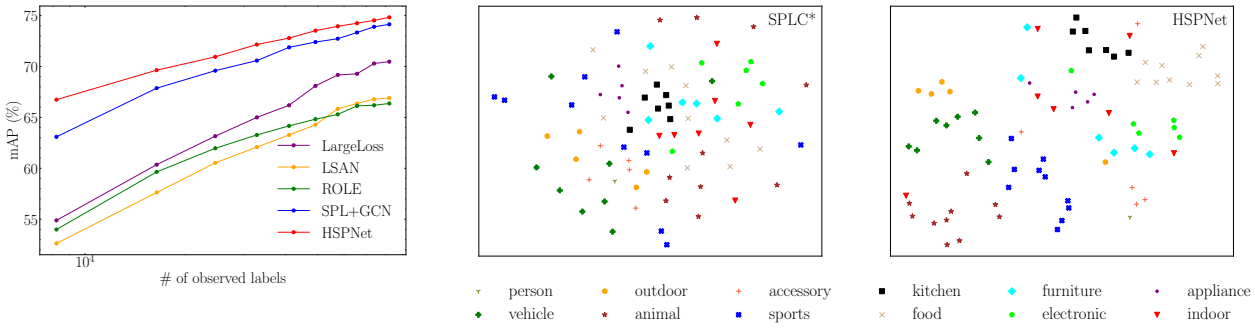

Figure 3: Left: Comparison results in the few-shot setting. Middle and right: Visualization results of label features learned by SPLC* and our HSPNet, respectively. Each color stands for one group. Labels in the same group are in the same color. Best viewed in color.

Table 4: Analyses on the automated hierarchical dependencies (%).

SPL	SPL+GCN	HSPNet-cluster	HSPNet
74.36	75.12	75.62	**75.71**

Table 5: Analyses on the learnable group-aware label prompt (%).

group-unaware	group-aware		
$\boldsymbol{w}_i$	$[\boldsymbol{s}_i, \boldsymbol{w}_i]$	$\boldsymbol{s}_i + \boldsymbol{w}_i$	$\boldsymbol{s}_i$
74.36	75.17	75.11	**75.18**

cluster-centric features as the group representations, i.e., $\boldsymbol{G}$ (see Equation (2)). As a result, we can automatically construct the hierarchical dependencies for SPML. We conduct experiments with the HSPNet, and denote HSPNet with such automated hierarchical dependencies as "HSPNet-cluster". As shown in Table 4, "HSPNet-cluster" can significantly outperform the baseline methods, i.e., SPL and SPL+GCN. Besides, it achieves comparable performance with our HSPNet. These results indicate that grouping labels according to the feature cluster is sufficient to provide rich hierarchical dependencies among labels and groups, and thus applicable when no explicit group information is provided.

Analyses on the group-aware label prompt. Here, we investigate the effect of different strategies to construct the prompt for labels with the hierarchical semantic structure in HCP. Given a label y_i with its embedding $\boldsymbol{e}_i$, we denote the label prompt as X_i which is fed into the text encoder $E(\cdot)$ with $\boldsymbol{e}_i$ (see Equation (4)). For ease of explanation, we denote the learnable label prompt for y_i in the single prompt learning (SPL) strategy like CoOp [44] as $\boldsymbol{w}_i = \{\boldsymbol{u}_{i1}, \boldsymbol{u}_{i2}, \boldsymbol{u}_{i3}, ..., \boldsymbol{u}_{iN}\}$. Given $\boldsymbol{s}_i = \{\boldsymbol{v}_{i1}, \boldsymbol{v}_{i2}, ..., \boldsymbol{v}_{iN}\}$ derived by Equation (3), we discuss four variants to obtain the prompt for label y_i: 1) directly using $\boldsymbol{s}_i$ like HCP, i.e., $X_i = \boldsymbol{s}_i$; 2) concatenating $\boldsymbol{s}_i$ and $\boldsymbol{w}_i$, i.e., $X_i = [\boldsymbol{s}_i, \boldsymbol{w}_i]$; 3) adding $\boldsymbol{s}_i$ with $\boldsymbol{w}_i$, i.e., $X_i = \boldsymbol{s}_i + \boldsymbol{w}_i$; 4) directly using $\boldsymbol{w}_i$ like SPL, i.e., $X_i = \boldsymbol{w}_i$. As shown in Table 5, three ways that construct group-aware prompts, i.e., $\boldsymbol{s}_i$, $[\boldsymbol{s}_i, \boldsymbol{w}_i]$ and $\boldsymbol{s}_i + \boldsymbol{w}_i$, can consistently outperform the single prompt, i.e., $\boldsymbol{w}_i$, with a similar margin, indicating the general advantage of the proposed hierarchical prompt learning. Considering that, compared with directly using $\boldsymbol{s}_i$, both $[\boldsymbol{s}_i, \boldsymbol{w}_i]$ and $\boldsymbol{s}_i + \boldsymbol{w}_i$ need to tune more prompt tokens due to the introduction of $\boldsymbol{w}_i$, we adopt $\boldsymbol{s}_i$ only in our HCP and HSPNet by default.

Table 6: Analyses on the HGCN (%).

HCP	w/o inter-group	w/o inter-label	HSPNet
75.18	75.52	75.39	**75.71**

Analyses on the hierarchical graph convolutional network. On top of HCP, we further verify the positive effects of inter-label and inter-group relationships used in HGCN. To achieve this goal, we wipe out the group-level GCN and the label-level GCN in the HGCN, separately, which are denoted as "w/o inter-group" and "w/o inter-label", respectively. As shown in Table 6, both inter-label and inter-group relationships can achieve consistent performance gains. Combining both kinds of relationships can further enhance the label hypothesis prediction, leading to an overall performance gain of 0.53% over HCP. These results demonstrate the effectiveness of exploring both kinds of relationships.

Results on the few-shot SPML setting. To investigate the performance of the proposed HSPNet with few training images in SPML, we conduct the experiments on the few-shot SPML setting, following [13]. We randomly sample training images with a ratio ranging from 10% to 100% as in [13]. As shown in Figure 3, our method can consistently outperform the state-of-the-art methods with a significant margin on different ratios of labels. Besides, as the number of labels decreases, the performance gains of the proposed HSPNet increase. Particularly, compared with LargeLoss, our method can obtain an improvement of 11.86% in terms of mAP when only given 10% of the training images. Compared with the strong baseline, i.e., SPL+GCN, the proposed HSPNet can still enjoy 1.38% mAP improvement on average. These experimental results demonstrate the effectiveness of the proposed method even with few training images in SPML. They also show the robustness of our method.

Table 7: Quantitative results compared with the state-of-the-art methods under the partial label setting (%). The bold and the underline indicate the best and the second best results, respectively. * indicates results that use CLIP weights as ours.

Datasets	Method	10%	20%	30%	40%	50%	60%	70%	80%	90%	Avg.
COCO	SSGRL	62.5	70.5	73.2	74.5	76.3	76.5	77.1	77.9	78.4	74.1
	GCN-ML	63.8	70.9	72.8	74.0	76.7	77.1	77.3	78.3	78.6	74.4
	SST	68.1	73.5	75.9	77.3	78.1	78.9	79.2	79.6	79.9	76.7
	SARB	71.2	75.0	77.1	78.3	78.9	79.6	79.8	80.5	80.5	77.9
	DualCoop*	**78.9**	81.1	81.9	82.7	82.9	83.2	83.4	83.7	83.9	82.4
	HSPNet (ours)	78.3	**81.4**	**82.2**	**83.6**	**84.3**	**84.8**	**85.0**	**85.4**	**85.6**	**83.4**
VG-200	SSGRL	34.6	37.3	39.2	40.1	40.4	41.0	41.3	41.6	42.1	39.7
	GCN-ML	32.0	37.8	38.8	39.1	39.6	40.0	41.9	42.3	42.5	39.3
	SST	38.8	39.4	41.1	41.8	42.7	42.9	43.0	43.2	43.5	41.8
	SARB	41.4	44.0	44.8	45.5	46.6	47.5	47.8	48.0	48.2	46.0
	DualCoop*	**45.6**	**47.6**	48.5	49.1	49.4	49.3	49.9	50.1	50.3	48.9
	HSPNet (ours)	43.0	46.7	**48.7**	**49.9**	**50.5**	**51.1**	**51.4**	**51.6**	**51.9**	**49.4**

Table 8: Quantitative results for the effect of different modules in our proposed HSPNet under the partial setting (%).

Model	COCO	VG-200	Avg.
SPL	81.54	47.55	64.55
SPL+GCN	82.12	48.11	65.12
HCP	82.60	48.43	65.52
HCP+HGCN (i.e., HSPNet)	**83.39**	**49.41**	**66.40**

Visualization analysis. To qualitatively show the effectiveness of our proposed method, we visualize the learned label features for COCO. We choose SPLC* as our baseline since it is a state-of-the-art method with CLIP weights as ours but without consideration of the label relationship. As shown in Figure 3, we can observe that, compared with SPLC*, the proposed HSPNet can generate more compact label feature clusters for SPML. Such a clustering result is consistent with the inherent hierarchical semantic structure among labels, indicating the success of the proposed method in learning better semantically discriminative label features and thus benefiting the image-label matching.

4.5 Results on Partial Label Setting

Details. To verify the generalization of the HSPNet to other practical MLC scenarios, we conduct experiments on standard benchmark datasets, i.e., COCO and Visual Genome (VG-200) [14], under the partial label setting, following [22, 27]. We randomly maintain partial annotations for the training images with a ratio ranging from 10% to 90% and the average results are reported. Please refer to Appendix A for details of experiments.

Results. As shown in Table 7, our method can also obtain superior results to the state-of-the-art methods in the partial label setting. Specifically, on the COCO dataset, compared with DualCoop which also leverages CLIP, the proposed HSPNet can achieve 1.0% mAP gain on average. On the VG-200 dataset, which is more difficult than COCO, our method can consistently obtain better results with gains of 0.5% mAP on average. Note that although our method is inferior to DualCoop under small ratios of partial labels, when the ratio is increasing, our method can quickly catch up DualCoop and significantly outperform it with maximum gains of 1.7% mAP (90% in COCO). These results adequately verify that our method can obtain consistent performance improvements under the partial label setting, well demonstrating its good generalization capability.

We further conduct the ablation study to verify the effectiveness of each modules under the partial setting. As shown in Table 8, each module can consistently achieve performance improvements on all benchmark datasets. Specifically, compared with SPL, the designed HCP can obtain an overall 0.97% mAP improvement. Based on HCP, our HGCN can further lead to a 0.88% mAP gain on average. Moreover, our proposed HSPNet can significantly outperform the SPL+GCN model by an overall 1.28% mAP, well demonstrating the superiority of our proposed method.

5 CONCLUSION

In this paper, we propose a hierarchical prompt learning method with a novel the Hierarchical Semantic Prompt Network (HSPNet) for single positive multi-label learning. We use the CLIP model to implement a hierarchical conditional prompt (HCP) strategy that can recognize the label-group dependency between labels and their corresponding groups. With the hierarchical graph convolutional network (HGCN), the inter-label and inter-group dependencies can be explored, enhancing the image-label matching. Thanks to both HCP and HGCN, the HSPNet can successfully harness the hierarchical semantic relationship among labels, thus leading to superior performance. Experimental results on multiple benchmark datasets show that our method can consistently outperform various baseline methods, well demonstrating its effectiveness and superiority.

6 ACKNOWLEDGMENTS

This research was supported by National Natural Science Foundation of China (Nos 62271281, 61925107, U1936202, 62021002) and Zhejiang Provincial Natural Science Foundation of China (No. LDT23F01013F01).

REFERENCES

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In European conference on computer vision. Springer, 213–229.

[2] Tianshui Chen, Tao Pu, Hefeng Wu, Yuan Xie, and Liang Lin. 2022. Structured semantic transfer for multi-label recognition with partial labels. In Proceedings of the AAAI conference on artificial intelligence, Vol. 36. 339–346.

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

[4] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. Nus-wide: a real-world web image database from national university of singapore. In Proceedings of the ACM international conference on image and video retrieval. 1–9.

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

[6] Xianzhi Du, Tsung-Yi Lin, Pengchong Jin, Golnaz Ghiasi, Mingxing Tan, Yin Cui, Quoc V Le, and Xiaodan Song. 2020. Spinenet: Learning scale-permuted backbone for recognition and localization. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 11592–11601.

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

[8] Mark Everingham and John Winn. 2012. The PASCAL visual object classes challenge 2012 (VOC2012) development kit. Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep 2007 (2012), 1–45.

[9] Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. arXiv preprint arXiv:2012.15723 (2020).

[10] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? Transactions of the Association for Computational Linguistics 8 (2020), 423–438.

[11] Warren Jouanneau, Aurélie Bugeau, Marc Palyart, Nicolas Papadakis, and Laurent Vézard. 2022. A patch-based architecture for multi-label classification from single label annotations. arXiv preprint arXiv:2209.06530 (2022).

[12] Bo Ke, Yunquan Zhu, Mengtian Li, Xiujun Shu, Ruizhi Qiao, and Bo Ren. 2022. Hyperspherical Learning in Multi-Label Classification. In European Conference on Computer Vision. Springer, 38–55.

[13] Youngwook Kim, Jae Myung Kim, Zeynep Akata, and Jungwoo Lee. 2022. Large Loss Matters in Weakly Supervised Multi-Label Classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 14156–14165.

[14] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. International Journal of Computer Vision (2017).

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023).

[16] Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, and Tao Mei. 2019. Vrr-vg: Refocusing visually-relevant relationships. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 10403–10412.

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

[18] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. 2021. Query2label: A simple transformer way to multi-label classification. arXiv preprint arXiv:2107.10834 (2021).

[19] Chaofan Ma, Yuhuan Yang, Yanfeng Wang, Ya Zhang, and Weidi Xie. 2022. Open-vocabulary Semantic Segmentation with Frozen Vision-Language Models. arXiv preprint arXiv:2210.15138 (2022).

[20] George A Miller. 1995. WordNet: a lexical database for English. Commun. ACM 38, 11 (1995), 39–41.

[21] Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. arXiv preprint arXiv:2111.09734 (2021).

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

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

[24] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric loss for multi-label classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 82–91.

[25] Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben-Baruch, and Asaf Noy. 2023. Ml-decoder: Scalable and versatile classification head. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 32–41.

[26] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. arXiv preprint arXiv:2010.15980 (2020).

[27] Ximeng Sun, Ping Hu, and Kate Saenko. 2022. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. arXiv preprint arXiv:2206.09541 (2022).

[28] Thomas Verelst, Paul K Rubenstein, Marcin Eichner, Tinne Tuytelaars, and Maxim Berman. 2023. Spatial consistency loss for training multi-label classifiers from single-label annotations. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 3879–3889.

[29] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. (2011).

[30] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. 2016. Cnn-rnn: A unified framework for multi-label image classification. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2285–2294.

[31] Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. 2020. Multi-label classification with label graph superimposing. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34. 12265–12272.

[32] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. arXiv preprint arXiv:2108.10904 (2021).

[33] Max Welling and Thomas N Kipf. 2016. Semi-supervised classification with graph convolutional networks. In J. International Conference on Learning Representations (ICLR 2017).

[34] Xiangping Wu, Qingcai Chen, Wei Li, Yulun Xiao, and Baotian Hu. 2020. AdaHGNN: Adaptive hypergraph neural networks for multi-label image classification. In Proceedings of the 28th ACM International Conference on Multimedia. 284–293.

[35] Ming-Kun Xie, Jiahao Xiao, and Sheng-Jun Huang. 2022. Label-Aware Global Consistency for Multi-Label Learning with Single Positive Labels. Advances in Neural Information Processing Systems 35 (2022), 18430–18441.

[36] Ning Xu, Congyu Qiao, Jiaqi Lv, Xin Geng, and Min-Ling Zhang. 2022. One Positive Label is Sufficient: Single-Positive Multi-Label Learning with Label Enhancement. arXiv preprint arXiv:2206.00517 (2022).

[37] Shichao Xu, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Zhu Qi. 2022. A Dual Modality Approach For (Zero-Shot) Multi-Label Classification. arXiv preprint arXiv:2208.09562 (2022).

[38] Vacit Oguz Yazici, Abel Gonzalez-Garcia, Arnau Ramisa, Bartlomiej Twardowski, and Joost van de Weijer. 2020. Orderless recurrent models for multi-label classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 13440–13449.

[39] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022).

[40] Jiawei Zhan, Jun Liu, Wei Tang, Guannan Jiang, Xi Wang, Bin-Bin Gao, Tianliang Zhang, Wenlong Wu, Wei Zhang, Chengjie Wang, et al. 2022. Global Meets Local: Effective Multi-Label Image Classification via Category-Aware Weak Supervision. In Proceedings of the 30th ACM International Conference on Multimedia. 6318–6326.

[41] Youcai Zhang, Yuhao Cheng, Xinyu Huang, Fei Wen, Rui Feng, Yaqian Li, and Yandong Guo. 2021. Simple and Robust Loss Design for Multi-Label Learning with Missing Labels. arXiv preprint arXiv:2112.07368 (2021).

[42] Donghao Zhou, Pengfei Chen, Qiong Wang, Guangyong Chen, and Pheng-Ann Heng. 2022. Acknowledging the Unknown for Multi-label Learning with Single Positive Labels. arXiv preprint arXiv:2203.16219 (2022).

[43] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 16816–16825.

[44] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. International Journal of Computer Vision 130, 9 (2022), 2337–2348.

[45] Yue Zhu, James T Kwok, and Zhi-Hua Zhou. 2017. Multi-label learning with global and local label correlation. IEEE Transactions on Knowledge and Data Engineering 30, 6 (2017), 1081–1094.

A EXPERIMENT SETTINGS

A.1 Datasets

For the single positive multi-label learning (SPML), we adopt four benchmark datasets, i.e., COCO [17], PASCAL VOC 2012 (VOC) [8],

Table 9: The statistics of all benchmark datasets.

Statistic	LargeLoss setup				SPLC setup				Partial label	
	COCO	VOC	NUS	CUB	COCO	VOC	NUS	CUB	COCO	VG-200
Classes	80	20	81	312	80	20	81	312	80	200
Groups	12	4	10	20	12	4	10	20	12	13
Training samples	65,665	4,574	120,000	4,795	82,081	5,717	119,103	5,994	82,081	82,904
Validation samples	16,416	1,143	30,000	1,199	-	-	-	-	-	-
Test samples	40,137	5,823	60,260	5,794	40,137	5,823	50,720	5,794	40,137	10,000
Avg.label/img	2.9	1.5	1.9	31.5	2.9	1.5	2.4	31.5	2.9	10.7

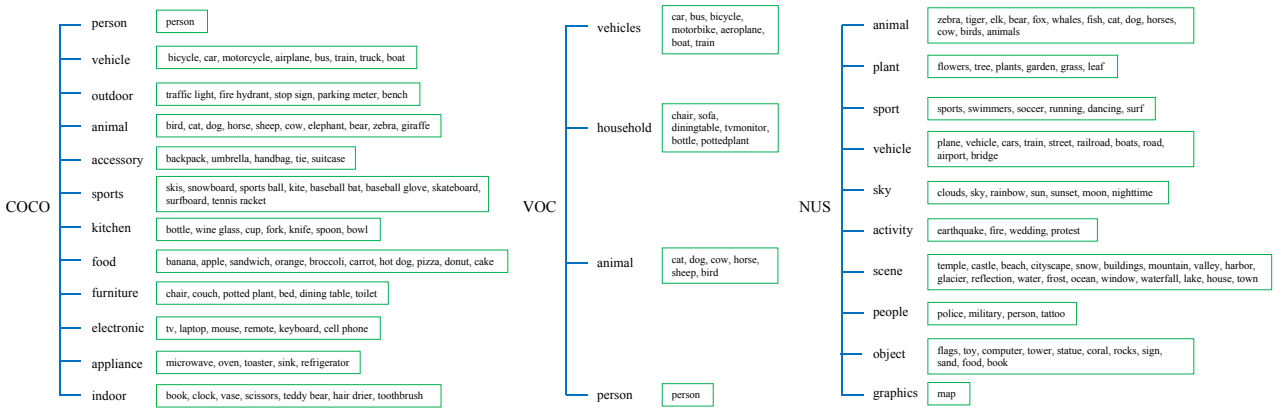

Figure 4: The hierarchical semantic structures of benchmark datasets in SPML (left: for COCO, middle: for VOC, right: for NUS).

NUSWIDE (NUS) [4], and CUB [29], following [5, 13, 41, 42]. Table 9 provides their statistics. The COCO dataset is one of the most popular large-scale labeled image datasets, widely used for various vision tasks, such as object detection [1], image captioning [21], instance segmentation [6] and classification [25], etc. It consists of 82,081 images in the training set and 40,137 images in the test set, covering 80 classes that belong to 12 super categories. In the LargeLoss setup, we split the training set into 65,665 images for training and 16,416 images for validation. The VOC dataset is a standardized image dataset for object recognition with a small label set of 20 categories. It contains 5,717 training images and 5,823 test images. We adopt 4,574 images for training and 1,143 images for validation in the LargeLoss setup. The NUS dataset is a large-scale web image dataset with 81 concepts. It is designed to explore web image annotation and retrieval problems [4]. Since not all images are available, we find two variants of NUS dataset in previous works, i.e., [13] and [41], which are used in the LargeLoss setup and SPLC setup, respectively. In the LargeLoss setup, NUS contains 120,000 images for training, 30,000 images for validation, and 60,260 images for test. In the SPLC setup, NUS consists of 119,103 training images and 50,720 test images. The CUB dataset is a fine-grained dataset for birds, covering 200 categories and 312 attributes. It contains 5,994 images in the training set and 5,794 images in the test set. In the LargeLoss setup, 4,795 images are used for training and 1,199 images are used for validation.

For the partial label setting, we conduct experiments on two benchmarks, i.e., COCO [17] and Visual Genome (VG-200) [14], following [2, 22]. The COCO dataset is identical to the one used in SPML. The Visual Genome dataset is a large-scale image-based knowledge base, widely used in cognitive vision tasks, e.g., visual relationship detection [16]. Following [22], the VG-200 dataset contains 200 frequent labels from the original Visual Genome dataset and the training set is composed of randomly selected 82,904 images while the left 10,000 images are in the test set.

A.2 Implementation details

For SPML, we adopt the concept taxonomies provided in the COCO, VOC, and NUS datasets to derive the hierarchical semantic structures, as shown in Figure 4. For the CUB dataset, we divide the classes into groups based on the part of the bird described by the attribute. Figure 5 provides the label-group correspondence of the CUB dataset. We train the SPML model for 30 epochs in total.

For the partial label setting, we design the hierarchical structure of the VG-200 dataset based on labels' semantic relationships in the WordNet [20], as shown in Figure 6. We adopt the same label-group correspondence for the COCO dataset as in SPML. For a fair comparison, we adopt the ResNet101-based CLIP, using the same image resolution 448×448. For data augmentation, we employ the widely used random horizontal flip and random resized crop, following [2, 22]. We use two GPUs with a batch size of 32 during training. Besides, we adopt the Adam optimizer and OneCycle learning rate schedule with the maximal learning rate of $1e^{-4}$. We also train the model for a total of 30 epochs.

Figure 5: The hierarchical semantic structure of the CUB dataset.

Figure 6: The hierarchical semantic structure of the VG-200 dataset.