Texts as Images in Prompt Tuning for Multi-Label Image Recognition

Zixian Guo[1,2*] Bowen Dong[1] Zhilong Ji[2] Jinfeng Bai[2] Yiwen Guo[4] Wangmeng Zuo[1,2(✉)]
[1]Harbin Institute of Technology [2]Tomorrow Advancing Life [3]Pazhou Lab, Guangzhou [4]Independent Researcher
zixian_guo@foxmail.com cndongsky@gmail.com zhilongji@hotmail.com
jfbai.bit@gmail.com guoyiwen89@gmail.com wmzuo@hit.edu.cn

Abstract

*Prompt tuning has been employed as an efficient way to adapt large vision-language pre-trained models (*e.g. *CLIP) to various downstream tasks in data-limited or label-limited settings. Nonetheless, visual data (*e.g., *images) is by default prerequisite for learning prompts in existing methods. In this work, we advocate that the effectiveness of image-text contrastive learning in aligning the two modalities (for training CLIP) further makes it feasible to treat texts as images for prompt tuning and introduce TaI prompting. In contrast to the visual data, text descriptions are easy to collect, and their class labels can be directly derived. Particularly, we apply TaI prompting to multi-label image recognition, where sentences in the wild serve as alternatives to images for prompt tuning. Moreover, with TaI, double-grained prompt tuning (TaI-DPT) is further presented to extract both coarse-grained and fine-grained embeddings for enhancing the multi-label recognition performance. Experimental results show that our proposed TaI-DPT outperforms zero-shot CLIP by a large margin on multiple benchmarks,* e.g., *MS-COCO, VOC2007, and NUS-WIDE, while it can be combined with existing methods of prompting from images to improve recognition performance further. The code is released at https://github.com/guozix/TaI-DPT.*

1. Introduction

Recent few years have witnessed rapid progress in large vision-language (VL) pre-trained models [1, 16, 19, 24, 33, 36] as well as their remarkable performance on downstream vision tasks. A VL pre-trained model generally involves data encoders, and it is becoming increasingly popular to exploit image-test contrastive loss [24] to align the embedding of images and texts into a shared space. When adapting to downstream tasks in data-limited or label-limited settings, it is often ineffective to fine-tune the entire model, due to its high complexity. Then, prompt tuning as a representative parameter-efficient learning paradigm has emerged as an efficient way to adapt VL models to downstream tasks.

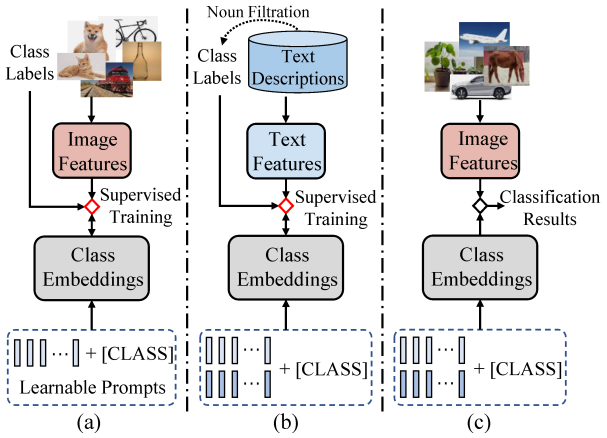

Figure 1. A comparison between prompting from images and our text-as-image (TaI) prompting. (a) Prompting from images (*e.g.*, [41]) uses labeled images of task categories to learn the text prompts. Instead, (b) our TaI prompting learn the prompts with easily-accessed text descriptions containing target categories. (c) After training, the learned prompts in (a) or (b) can be readily applied to test images.

Albeit considerable achievements have been made, existing prompt tuning methods generally require visual data to learn prompts (as shown in Fig. 1(a)). For example, CoOp [41] learns from annotated images. CoCoOp [40] further introduces generalizable input-conditional prompts. DualCoOp [28] adapts CLIP to multi-label recognition tasks by training pairs of positive and negative prompts with partial-labeled images. Nonetheless, the performance of these prompting methods may be limited when it is infeasible to obtain sufficient image data or annotate the required images.

In this paper, we advocate treating **T**exts **a**s **I**mages for prompt tuning, *i.e.*, TaI prompting. It is considered feasible as the image encoder and text encoder in many pre-trained VL models [16, 24] encode images and texts into a shared space. Given an image and its caption, the visual features produced by the image encoder will be close to the text feature of the caption produced by the text encoder. There-

*This work was done when Zixian Guo was a research intern at TAL.

fore, in addition to extracting visual features from images, it is also feasible to extract text features as alternatives form, for example, descriptive sentences and captions, for prompt tuning (see Fig. 1(b)). TaI prompting has several interesting properties and merits. Taking a downstream image recognition task as an example, given a set of object categories, one can easily crawl a large set of text descriptions that contain object names from these categories. Text descriptions are easily accessible in this way, and class labels can be directly derived from text descriptions, which means, in contrast to prompting from images, TaI prompting may suffer less from the data-limited and label-limited issues.

We use multi-label image recognition [8, 9, 11, 20, 35] to verify the effectiveness of our TaI prompting in this paper. To begin with, we crawl the captions from the public image caption datasets (*e.g.*, MS-COCO [20]) and localized narratives from object detection datasets (*e.g.*, Open Images [18]) to form the training set of text descriptions. For any specific multi-label recognition task, we adopt a noun filter to map the nouns in the text descriptions to the corresponding object categories, and then only keep the text descriptions that contain one or more classes of target objects. To better cope with multi-label classification, we introduce double-grained prompt tuning (*i.e.*, TaI-DPT) which involves: (i) a set of global prompts to generate embeddings for classifying whole sentences or images, and (ii) a set of local prompts to extract embeddings for discriminating text tokens or image patches. Given a set of text descriptions, global and local prompts can be tuned by minimizing the ranking loss [14]. Note that, though these prompts are learned from text descriptions solely, they can be readily deployed to classify whole images as well as image patches during testing (see Fig. 1(c)). Experimental results show that, without using any labeled images, our TaI prompting surpasses zero-shot CLIP [24] by a large margin on multiple benchmarks, *e.g.*, MS-COCO, VOC2007, and NUS-WIDE.

Moreover, when images are also available during training, our TaI prompting can be combined with existing methods of prompting from images to improve its performance. In particular, given a few annotated images, our TaI-DPT can be integrated with CoOp as a prompt ensemble for improving classification accuracy. With partially labeled training data being provided, we may also combine TaI-DPT and DualCoOp [28] to improve multi-label recognition accuracy consistently. Extensive results verify the effectiveness of our TaI-DPT in comparison to state-of-the-art. To sum up, the contributions of this work include:

- We propose Texts as Images in prompt tuning (*i.e.*, TaI prompting) to adapt VL pre-trained models to multi-label image recognition. Text descriptions are easily accessible and, in contrast to images, their class labels can be directly derived, making our TaI prompting very compelling in practice.
- We present double-grained prompt tuning (*i.e.* TaI-DPT) to extract both coarse-grained and fine-grained embeddings for enhancing multi-label image recognition. Experiments on multiple benchmarks show that TaI-DPT achieves comparable multi-label recognition accuracy against state-of-the-arts.
- The prompts learned by TaI-DPT can be easily combined with existing methods of prompting from images in an off-the-shelf manner, further improving multi-label recognition performance.

2. Related Work

2.1. Multi-Label Image Recognition

Multi-label image recognition [3,7,8,12,15,21,32,35,37] aims to recognize all the object categories [11, 20] or concepts [9] in an input image. Various modules [6, 30] have been introduced to better represent the inter-class relationships, and modern classification losses [3, 14] have been used to make model learning easier.

To model the label dependencies, CNN-RNN [30] introduces recurrent neural networks, *e.g.*, RNN and LSTM, to predict appeared classes in a sequential manner. [6,8,31,35] use graph convolution modules to learn the correlation between class labels. CHAMP [29] measures the severity of misclassification by building a domain-specific hierarchy tree according to the relation of categories, where each class is related to a tree node, to improve the robustness of the model. Albeit effective, these methods require a considerable number of labeled images to learn the category relationships sufficiently. While in data-limited or label-limited regimes, *e.g.*, few-shot or partial-label data, it will be difficult for these models to learn well as expected. Specifically designed loss functions also struggle to obtain significant improvements when learning with limited data.

Multi-Label Recognition from Few-shot Samples. To better exploit the small number of samples, LaSO [2] synthesizes samples by manipulating the features of paired training images. Different ways of manipulating label sets are used to train the model, resulting in generalizable discriminative features. [27] introduces a meta-learning framework for better learning of past tasks and generalization to new tasks, and leverages the number of labels as useful information for learning.

Multi-Label Recognition from Partial-label Data. Partial-label refers to the scenarios where some labels are unknown. [10] propose a normalized BCE loss to balance the proportion of known labels. [5] learns to complement unknown labels by utilizing within-image and cross-image semantic correlations. [23] blends the representation of training images and class proxies to compensate for the loss of information due to unknown labels.

Albeit significant progress has been made, it remains a challenging issue for learning multi-label image recognition

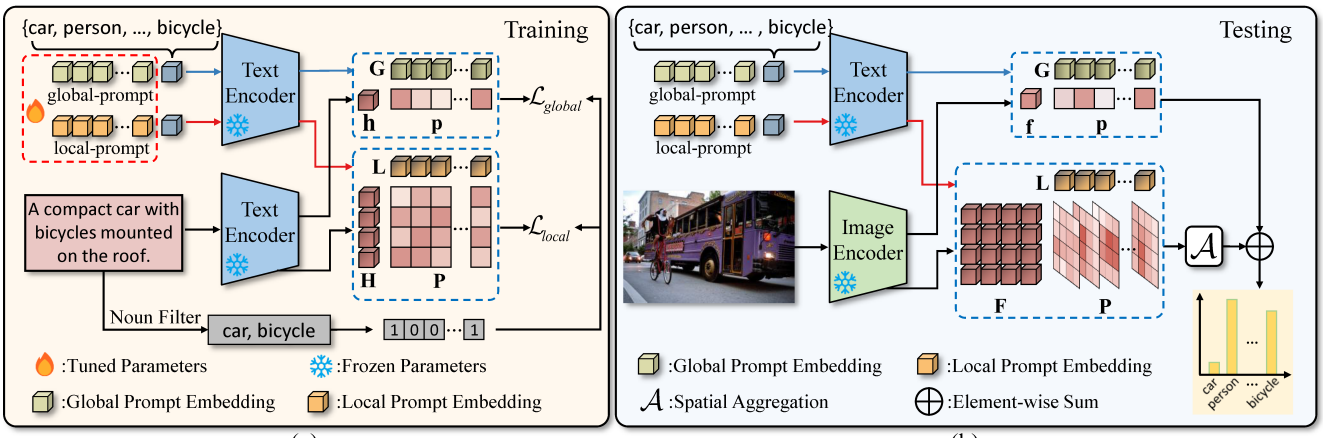

Figure 2. Training and testing pipeline of our proposed Text-as-Image (TaI) prompting, where we use text descriptions instead of labeled images to train the prompts. (a) During training, we use two identical text encoders from pre-trained CLIP to extract the global & local class embeddings ($\mathbf{G}$&$\mathbf{L}$) and overall & sequential text embeddings ($\mathbf{h}$&$\mathbf{H}$) respectively from the prompts and text description. The corresponding cosine similarity ($\mathbf{p}$&$\mathbf{P}$) between the embeddings are guided by the derived pseudo labels with ranking loss. (b) During testing, we replace the input from text descriptions to images. The global and local class embeddings can discriminate target classes from global & local image features ($\mathbf{f}$&$\mathbf{F}$). The final classification results are obtained by merging the scores of the two branches.

in image-limited or label-limited regimes. Built upon VL pre-trained models, this paper suggests learning prompts from texts instead of images, thereby offering a novel yet complementary perspective for handling low-resource multi-label image recognition.

2.2. Prompt Tuning for Vision-Language Models

To transfer pre-trained knowledge to downstream tasks in data-limited settings, prompt tuning [13,17,34,39,41,42] has become a popular parameter-efficient way to achieve the goal, due to its flexibility and ease of use. CoOp [41] learns the prompts by using (a few) annotated images of each class from the target dataset. CoCoOp [40] further proposes to improve CoOp [41] by formulating the prompts in an image-conditional way to maintain better generalization to unseen classes. To avoid overfitting, ProGrad [42] leverages predictions from zero-shot CLIP to regularize gradients in prompt learning process. TPT [26] suggests optimizing test-time prompts by promoting the consistency of augmented test images. ProDA [22] uses multiple pieces of prompts to estimate the distribution of classifier weights for better handling of varying visual features. DualCoOp [28] firstly adapts CLIP to multi-label image recognition with partially labeled data by learning pairs of positive and negative prompts for each class to ensure independent binary classification for each class.

Albeit existing prompt tuning approaches have achieved significant improvements in downstream tasks, images as well as a portion of class labels are prerequisite to supervise the optimization of the learnable prompts. In this paper, we propose to treat texts as images in prompt tuning, which, compared to labeled images, are much easier to collect with existing caption datasets and modern search engines. Our proposed TaI-DPT surpasses zero-shot CLIP by a large margin and can be combined with the prompts learned by existing methods of prompting from images to further boost recognition performance.

3. Proposed Method

Here we present Text-as-Image prompting, *i.e.*, TaI prompting, for adapting pre-trained VL models to multi-label image recognition. Our TaI prompting uses only easily-accessed free-form texts as training data to learn effective prompts for downstream multi-label recognition tasks. To begin with, We present an overview of TaI prompting in Sec. 3.1. Then, we introduce our preparation of training texts in Sec. 3.2. We further explain the design of the double-grained prompt tuning (*i.e.*, TaI-DPT) in Sec. 3.3, and provide the loss function used to train the model in Sec. 3.4. Finally, we propose to ensemble prompts from TaI-DPT with prompts learned from images as a flexible integration to improve multi-label recognition performance.

3.1. Overview of Our Method

Fig. 2 illustrates the design of our proposed TaI-DPT framework, including the training and testing phases. During training, we learn prompts with only supervision from texts. Two identical copies of the text encoder $\mathrm{Enc_T}$ from the pre-trained CLIP are used to encode the prompts and text data, respectively. We introduce two sorts of trainable prompts (*i.e.*, the global prompts and local prompts) to obtain global and regional class embeddings. A noun filtering strategy is used to generate classification pseudo-labels for each text description, which is applied to supervise the classification scores obtained by calculating the cosine sim-

ilarity of class embeddings and text features. Only the parameters in prompts are optimized in the training phase, while the text encoders are both kept frozen. During testing, the class embeddings are obtained by encoding the two sets of learned prompts with the text encoder $\mathrm{Enc_T}$ as in training, while the other input source changes from text descriptions to test images. Pre-trained image encoder $\mathrm{Enc_I}$ from CLIP is used to extract global and dense features of each test image, then the cosine similarity are computed between features and class embeddings of the global and local prompts. The final classification result is obtained by fusing the global and local cosine scores. In the following, we explain the details of our proposed method.

3.2. Preparation of Text Descriptions

To obtain sufficient category information from the language that helps in image recognition, we have to ensure that: 1) the collected texts should contain content-rich descriptions of a scene, and 2) the contents of all text descriptions need to cover the category set of the target dataset for comprehensive representations of all categories. With the aim of ensuring reproducibility, we use readily available captions from the public image caption datasets (*e.g.*, MS-COCO [20]) and localized narratives from object detection datasets (*e.g.*, OpenImages [18]) as our language data source, while avoiding the workloads associated with randomly crawling texts from the Internet in this paper. Note that although each caption is paired with a corresponding image and human-annotated labels, we only use the captions, and no information from the pictures and labels is disclosed during training.

For a target multi-label recognition dataset $\mathcal{X}$ that has a category set $\mathcal{S} = \{\mathrm{s}_1, \mathrm{s}_2, \mathrm{s}_3, ..., \mathrm{s}_C\}$, where C denotes the number of categories and $\mathrm{s_i}$ denotes particular class name like "dog", "plane", etc., we search for sentences that contain at least one class name $\mathrm{s_i}$ in $\mathcal{S}$. Since multiple words or phrases usually exist to represent the same meaning for each class, searching solely for exact match of category names in texts may lead to many false negatives in the obtained pseudo ground-truth labels, which is harmful to prompt tuning. Towards tackling this issue, we introduce a noun filter to map nouns with similar meanings into the corresponding class label. Specifically, we construct a synonym dictionary $\mathcal{D}$ by including common synonyms of each class name in the target dataset. If a word in a text description matches any synonym of a specific class name, it is considered to contain a description of that category. Several examples of synonyms are shown as follows:

```
{'dog','pup','puppy','doggy'}
{'person','people','man','woman','human'}
{'bicycle','bike','cycle'}
{'car','taxi','automobile'}
{'boat','raft','dinghy'}
...
```

More details of the dictionary $\mathcal{D}$ are provided in the *Suppl*.

Then we conduct noun filtration by the following steps. First, for each text description, we use the tokenizer and lemmatizer from NLTK [4] to recover the stem of each word in the sentences. Next, for all keywords in $\mathcal{D}$, which contains all synonyms of the category set $\mathcal{S}$, we search in our language data source for sentences that contains at least one class name. For the text descriptions that do not match any synonym of any class name, we simply drop them away to ensure each piece of data has at least one concerned label. Finally, for each retained text description, we convert the class names it contains into binary pseudo-ground-truth vectors by setting classes that appear as positive and other classes as negative, following the order of class labels in the target dataset $\mathcal{X}$.

The word-level filtered labels may not be precisely correct since our searching strategy mentioned above is rather simple considering the diversity of free-form texts, where complex paraphrases and misspellings that widely exist in the corpus are not fully addressed. However, such a simple noun filtration can guarantee the reproducibility of this work, and our experiments also demonstrate that this simple and efficient data preparation leads to compelling results of multi-label recognition of our TaI prompting.

3.3. Text-as-Image for Dual-grained Prompt Tuning

Following [41], a prompt is defined as:

$$\boldsymbol{t}_i = [\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, ..., \boldsymbol{v}_M, \boldsymbol{s}_i] \tag{1}$$

where $i \in \{1, 2, ..., C\}$ is the class index, $\boldsymbol{s}_i$ denotes word embedding of the i-th class name $\mathrm{s_i}$. For $j \in \{1, \ldots, M\}$, $\boldsymbol{v}_j$ is a learnable word embedding whose dimension is the same as the dimension of normal word embeddings in the vocabulary. Just like in previous methods, *e.g.* CoOp [41], the prompts are learned by maximizing the probability of classifying each image into its ground-truth class:

$$p(y=i|\boldsymbol{x}) = \frac{\exp(\langle \mathrm{Enc_T}(\boldsymbol{t}_i), \mathrm{Enc_I}(\boldsymbol{x}) \rangle / \tau)}{\sum_{j=1}^{C} \exp(\langle \mathrm{Enc_T}(\boldsymbol{t}_j), \mathrm{Enc_I}(\boldsymbol{x}) \rangle / \tau)} \tag{2}$$

where $\boldsymbol{x}$ denotes the labeled training image and $\langle \cdot, \cdot \rangle$ calculates the cosine similarity.

After large-scale image-text contrastive pre-training, text features have been well-aligned to the image features of the same semantic meanings. Thus, features of texts that describe a certain type of object also exhibit categorical discriminability. Therefore, based on the aligned VL representation, we advocate considering the feature of a piece of text description that describes a specific category, as an alternative to an image feature.

Apart from using the global sentence representation (*i.e.*, the coarsest-grained text feature), we find that the sequential feature of word tokens from CLIP also possesses rich fine-grained information which is very similar to the region feature of dense image feature. In CLIP [24], cosine similarity

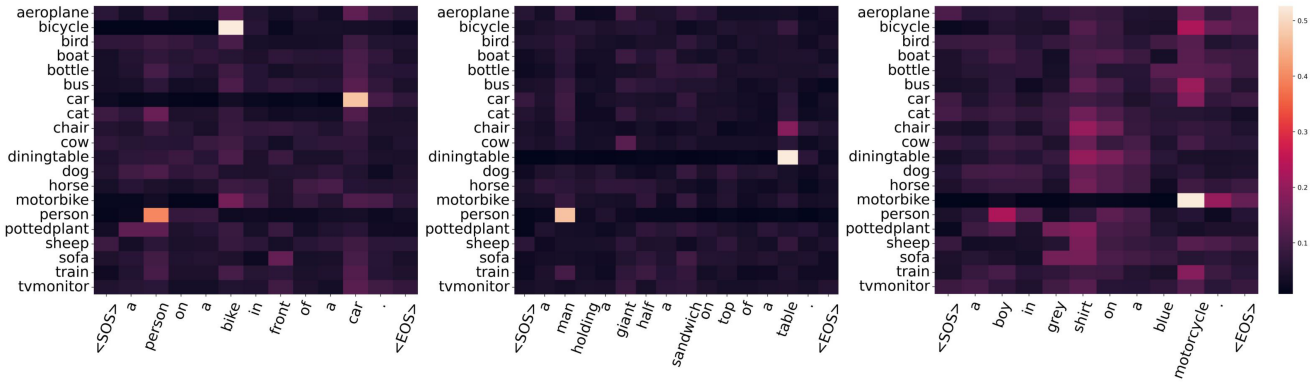

Figure 3. Visualization of correlations $\boldsymbol{P}$ between the local class embedding $\boldsymbol{L}$ and sequential token feature from texts. Each class embedding clearly correlates to words that describe the corresponding class (shown in highlight regions) rather than the global <EOS> token.

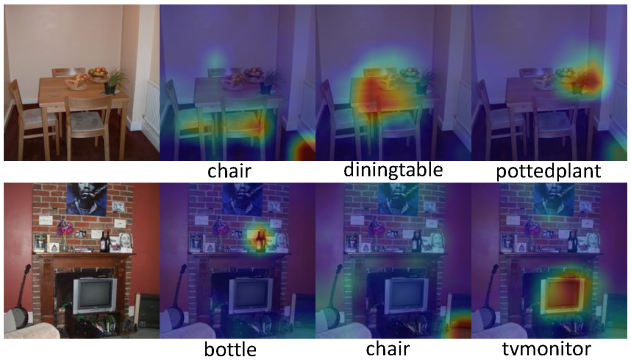

Figure 4. Visualization of correlations between the local class embedding $\boldsymbol{L}$ and dense image feature. The learned class embeddings can focus on the location of the object effectively.

between global image features, obtained by visual attention pooling, and global text features, obtained by projecting the feature of the last <EOS> token, are directly supervised with contrastive loss. In general, the global feature is sufficient for single-label classification because the target object usually is prominent in the picture. However, in multi-label recognition, the global feature is usually dominated by major objects, suppressing the recognition of non-significant objects concurrently existing in the image. Thus, it motivates us to explore fine-grained features and avoid the domination of the overly prominent object.

To grasp the discriminability of global features as well as learn from fine-grained features, we propose double-grained prompt tuning (*i.e.*, TaI-DPT) that uses two sets of prompts to handle global (*i.e.*, the coarsest-grained level) and local (*i.e.*, the fine-grained level) features, respectively, in two parallel branches. Formally, the double-grained prompt is defined as follows:

$$\begin{aligned} \boldsymbol{t}_i^G &= [\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, ..., \boldsymbol{v}_M, \boldsymbol{s}_i], \\ \boldsymbol{t}_i^L &= [\boldsymbol{v}'_1, \boldsymbol{v}'_2, \boldsymbol{v}'_3, ..., \boldsymbol{v}'_M, \boldsymbol{s}_i], \end{aligned} \tag{3}$$

where $\boldsymbol{v}_j$ and $\boldsymbol{v}'_j$ are learnable embeddings that are concatenated with word embedding $\boldsymbol{s}_i$ of the i-th class to obtain the global prompt $\boldsymbol{t}_i^G$ and local prompt $\boldsymbol{t}_i^L$, respectively. The sequences in Eq. (3) are fed to a copy of the text encoder $\mathrm{Enc_T}$ of CLIP to generate global and local class embeddings for each class, *i.e.* $\boldsymbol{G}_i = \mathrm{Enc_T}(\boldsymbol{t}_i^G)$ and $\boldsymbol{L}_i = \mathrm{Enc_T}(\boldsymbol{t}_i^L)$, $\boldsymbol{G} = \{\boldsymbol{G}_i\}_{i=1}^C$ and $\boldsymbol{L} = \{\boldsymbol{L}_i\}_{i=1}^C$ are encouraged to be correlated with global and local features, respectively. Note that the proposed double-grained prompts are different from dual prompts [28], which include a pair of contrastive positive and negative prompts for each class (More discussion about the differences between our method and DualCoOp is provided in the *Suppl*).

To preserve the fine-grained region features for the input image, we maintain the feature map before attention pooling layer of CLIP. As for the input text description, we preserve the sequential token features of the entire sentence instead of only the <EOS>token features. So we have:

$$\begin{aligned} \{\boldsymbol{f}, \boldsymbol{F}\} &= \mathrm{Enc_I}(\boldsymbol{x}), \\ \{\boldsymbol{h}, \boldsymbol{H}\} &= \mathrm{Enc_T}(\boldsymbol{r}), \end{aligned} \tag{4}$$

where $\boldsymbol{r}$ denotes a piece of training text description. $\boldsymbol{f}, \boldsymbol{h} \in \mathbb{R}^D$ are the extracted global image and text features. $\boldsymbol{F} \in \mathbb{R}^{N_1 \times D}$ and $\boldsymbol{H} \in \mathbb{R}^{N_2 \times D}$ are the flattened dense image features and sequential token features, respectively, where $N_1 = H \times W$ denotes the flattened spatial dimension of visual feature and N_2 denotes the length of text tokens.

Then, the global and local similarities are computed by:

$$\boldsymbol{p}_i = \langle \boldsymbol{u}, \boldsymbol{G}_i \rangle, \boldsymbol{P}_{ij} = \langle \boldsymbol{U}_j, \boldsymbol{L}_i \rangle \tag{5}$$

where $\boldsymbol{u}$ denotes either language feature $\boldsymbol{h}$ in training or visual feature $\boldsymbol{f}$ in testing, and $\boldsymbol{U}$ denotes $\boldsymbol{H}$ or $\boldsymbol{F}$ coordinately. Information in local branch $\boldsymbol{P}$ (visualized in Fig. 3 and Fig. 4) is aggregated in a spatially weighted manner:

$$\boldsymbol{p}'_i = \sum_{j=1}^N \frac{\exp(\boldsymbol{P}_{ij}/\tau_s)}{\sum_{j=1}^N \exp(\boldsymbol{P}_{ij}/\tau_s)} \cdot \boldsymbol{P}_{ij} \tag{6}$$

where τ_s accommodates the extent of focusing on a specific location. $\boldsymbol{p}_i$ and $\boldsymbol{p}'_i$ are optimized by the loss terms $\mathcal{L}_{global}$

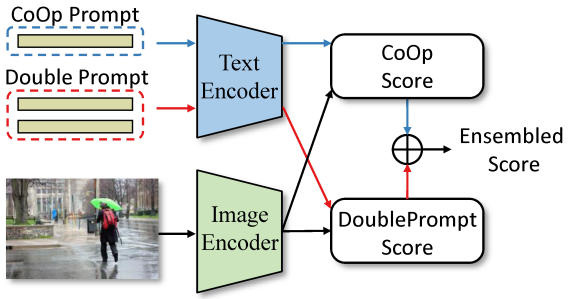

Figure 5. Our learned double-grained prompt tuning is easy to combine with existing prompt tuning methods with ensemble.

and $\mathcal{L}_{local}$, respectively, which we will discuss in Sec. 3.4. And in the testing phase, $\boldsymbol{p}$ and $\boldsymbol{p}'$ are combined to obtain the final classification score.

The visualization in Fig. 3 and Fig. 4 show that the learned local class embedding $\boldsymbol{L}$ can focus on the specific location where corresponding classes appear, both in sentences and images, even if the fine-grained visual and language features are not explicitly supervised in the training of CLIP.

3.4. Learning Objective

The overall learning objective is defined as $\mathcal{L} = \mathcal{L}_{global} + \mathcal{L}_{local}$, where $\mathcal{L}_{global}$ and $\mathcal{L}_{local}$ are loss terms for global text embedding and local text tokens, respectively. We adopt the ranking loss [14] to measure the discrepancy between classification scores and pseudo-ground-truth labels. Specifically, $\mathcal{L}_{global}$ and $\mathcal{L}_{local}$ are formulated as follows:

$$\begin{aligned} \mathcal{L}_{global} &= \sum\nolimits_{i\in\{c^+\}}\sum\nolimits_{j\in\{c^-\}} \max(0, m-\boldsymbol{p}_i+\boldsymbol{p}_j), \\ \mathcal{L}_{local} &= \sum\nolimits_{i\in\{c^+\}}\sum\nolimits_{j\in\{c^-\}} \max(0, m-\boldsymbol{p}'_i+\boldsymbol{p}'_j) \end{aligned} \tag{7}$$

where $\boldsymbol{p}$ and $\boldsymbol{p}'$ are global and aggregated local similarities described in Sec. 3.3, m is the margin controlling how much higher the similarity score with the positive classes is than with the negative classes. During training, we minimize the overall objective $\mathcal{L}$ with frozen text encoders, by optimizing the global and local prompts.

Binary cross-entropy loss and its variations such as asymmetric loss [25] have shown remarkable results in classification tasks [25, 28]. They are generally accompanied by a sigmoid function $\boldsymbol{\sigma}(x) = 1/(1+\exp(-x))$ to convert model outputs to probabilities. Nevertheless, we found that directly optimizing the probability $\boldsymbol{\sigma}(\boldsymbol{p})$ leads to a performance gap between training texts and testing images. We attribute this phenomenon to the modality gap between vision and language. Thus we consider ranking loss to be a more flexible and suitable way of supervision in the presence of the modality gap. A comparison of results between different loss functions is provided in *Suppl*.

3.5. Incorporating with Prompting from Images

Though our TaI-DPT is very different from existing methods of prompting from images, it is also complementary to them. To show this, we utilize an off-the-shelf prompt ensemble strategy to combine our TaI-DPT with existing methods in this section. As illustrated in Fig. 5, using CoOp [41] as an example, we can simply combine the scores of CoOp [41] and that of our TaI-DPT in a weighted sum manner. In particular, our TaI-DPT can be integrated with CoOp [41] when a few annotated images are provided and integrated with DualCoOp [28] when partially labeled training data are available.

We ensemble prompts by fusing the predicted scores, rather than averaging the class embeddings generated by different prompts, since the image encoder used in different methods may be different (*e.g.* we conduct our experiments with ResNet50, while DualCoOp uses ResNet101 for partial-label prompting). So ensembling with the classification score is more convenient. In Sec. 4.3, we also empirically show that our prompt ensemble strategy is effective in advancing multi-label recognition performance in the few-shot and partially labeled settings.

4. Experiments

4.1. Implementation Details

Architecture. We adopt CLIP ResNet-50 [24] as the visual encoder, and use the CLIP Transformer as the text encoder. During training, the parameters of the two encoders are kept frozen, and only learnable prompts are optimized.

Learnable Prompts. Our learnable prompts are shared among classes of all datasets. Class-specific prompting [41] (*i.e.*, an individual set of parameters for each category) has also been explored, but brings limited benefits. We initialize the value of each parameter with the Gaussian noise sampled from $\mathcal{N}(0, 0.02)$. The length of both the global prompts and local prompts are set to M = 16, while a longer sequence brings trivial improvements.

Datasets. To evaluate our TaI-DPT, we conduct the experiments on VOC2007 [11], MS-COCO [20], and NUS-WIDE [9]. VOC2007 contains 20 common categories, and following [6, 8, 28], we form the training/test set based on the official `trainval/test` split (5,011 images/4,952 images). MS-COCO includes 80 categories with 82,081 training images and 40,504 testing images. NUS-WIDE includes 81 concepts with 161,789 images for training. We adopt its test set (107,859 images) to evaluate our method. For zero-shot experiments in Sec. 4.2, the training sets of the datasets are not used, and we use only text data to learn the prompts as mentioned in Sec. 3.2. Besides, for VOC2007 and MS-COCO, the language data sources are captions from MS-COCO. For NUS-WIDE, we introduce localized narratives from OpenImages [18], which have a

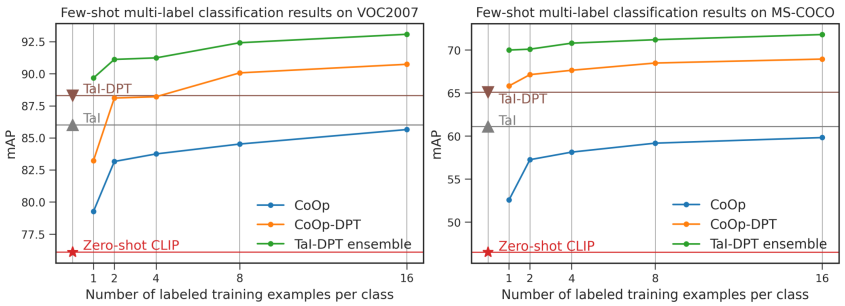

Figure 6. Comparison of different methods in few-shot multi-label recognition on VOC2007 and MS-COCO. Our zero-shot TaI-DPT can achieve comparable results with methods trained by 16-shot labeled image samples. And learned prompt ensemble proofs the complementarity between images and texts.

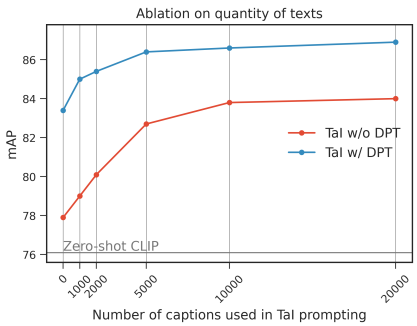

Figure 7. Ablation experiment on the number of texts and performance of TaI prompting on VOC2007.

Table 1. Comparison with zero-shot methods on VOC2007, MS-COCO, and NUS-WIDE. Our proposed TaI-DPT outperforms CLIP [24] by a large margin on all datasets.

Method	DPT	VOC2007	MS-COCO	NUSWIDE
ZSCLIP	✗	76.2	47.3	36.4
	✓	77.3	49.7	37.4
TaI	✗	86.0	61.1	44.9
	✓	**88.3**	**65.1**	**46.5**

broader range of content, to cover all the concepts in NUS-WIDE. In Sec. 4.3 and Sec. 4.4, for each dataset, the corresponding training data is used to conduct the experiments of partial-label and few-shot multi-label classification.

Training Details. We adopt SGD optimizer, and the training epoch is set to 20 for all datasets. The learning rates for MS-COCO, VOC2007, and NUS-WIDE are empirically initialized with 1e-4, 1e-4, and 1e-3, and decay by the cosine annealing rule. For ranking loss, we choose $m = 1$, and scale the $\boldsymbol{p}$ and $\boldsymbol{p}'$ by a factor of 4. τ_s is set as 0.02 via validation.

4.2. Comparison with Zero-Shot Methods

To demonstrate the effectiveness of our proposed TaI and DPT, we first compare it with the zero-shot CLIP (ZSCLIP). For fair comparison, we also introduce the DPT to ZSCLIP. Specifically, we adopt two identical default prompts "a photo of a [CLASS]" to separately deal with global and local features as DPT does. Table 1 lists the comparison results on VOC2007 [11], MS-COCO [20], and NUS-WIDE [9] datasets. From the table, our TaI prompting surpasses ZSCLIP by a large margin of 9.8%, 13.8%, and 8.5% mAP on VOC2007, MS-COCO, and NUS-WIDE, respectively, showing the effectiveness of our TaI. Furthermore, after training with fine-grained token features extracted from texts, our proposed DPT demonstrates a more powerful capability of discriminating local features than the default hand-crafted prompts and single global prompts.

Table 2. Comparison with existing multi-label few-shot learning methods on MS-COCO. The evaluation is based on mAP for zero-shot, 1-shot and 5-shot with 16 novel classes.

Method	**0-shot**	1-shot	5-shot
LaSO [2]	-	45.3	58.1
ML-FSL [27]	-	54.4	63.6
CoOp [41]	40.2 (ZSCLIP)	46.9	55.6
Tip-Adapter [38]	40.2 (ZSCLIP)	53.8	59.7
TaI-DPT	59.2	-	-

4.3. Comparison with Few-Shot Methods

We further compare with multi-label few-shot learning methods to verify the effectiveness of our TaI-DPT. In contrast to the well-studied single-label few-shot classification problem, few works tackle the multi-label few-shot scenario. Existing methods [2, 27] often deploy models trained on seen classes to few-shot novel classes. In Table 2, we compare zero-shot TaI-DPT to few-shot methods on 16 novel classes (see [2] for details about data split). TaI-DPT is comparable to the methods trained on 5-shot samples.

Besides, we consider a new multi-label few-shot setting where all the classes are regarded as novel classes. We select 1, 2, 4, 8, and 16-shot samples for each category following the strategy in [2]. For fair comparison, we train CoOp [41] and our TaI in the same settings, and we also extend them with DPT for a more comprehensive comparison. For CoOp-DPT, we set two sets of learnable prompts, to deal with global and local features, respectively. The results are illustrated in Fig. 6. One can see that, even without any image data regarding novel classes, our TaI can achieve comparable results to CoOp trained on 16-shot. Similar trends with the MS-COCO dataset and the DPT setting support our observation that the discriminative feature of text data can be used as images for prompting. Moreover, benefiting from the flexibility of prompts, we can easily integrate our TaI-DPT with CoOp-DPT by ensembling as said in Sec. 4.3. As illustrated in Fig. 6, though CoOp-DPT has

Table 3. Results of integrating our TaI-DPT with partial-label multi-label recognition method based on pre-trained CLIP. Our approach further improves the frontier performance of DualCoOp [28]. $*$ indicates the results based on our own reproduction.

Datasets	Method	10%	20%	30%	40%	50%	60%	70%	80%	90%	Avg.
MS-COCO	SARB [23]	71.2	75.0	77.1	78.3	78.9	79.6	79.8	80.5	80.5	77.9
	DualCoOp [28]	78.7	80.9	81.7	82.0	82.5	82.7	82.8	83.0	83.1	81.9
	DualCoOp*	81.0	82.3	82.9	83.4	83.5	83.9	84.0	84.1	84.3	83.3
	+TaI-DPT	**81.5**	**82.6**	**83.3**	**83.7**	**83.9**	**84.0**	**84.2**	**84.4**	**84.5**	**83.6**
PascalVOC 2007	SARB [23]	83.5	88.6	90.7	91.4	91.9	92.2	92.6	92.8	92.9	90.7
	DualCoOp [28]	90.3	92.2	92.8	93.3	93.6	93.9	94.0	94.1	94.2	93.2
	DualCoOp*	91.4	93.8	93.8	94.3	94.6	94.7	94.8	94.9	94.9	94.1
	+TaI-DPT	**93.3**	**94.6**	**94.8**	**94.9**	**95.1**	**95.0**	**95.1**	**95.3**	**95.5**	**94.8**
NUS-WIDE	DualCoOp*	54.0	56.2	56.9	57.4	57.9	57.9	57.6	58.2	58.8	57.2
	+TaI-DPT	**56.4**	**57.9**	**57.8**	**58.1**	**58.5**	**58.8**	**58.6**	**59.1**	**59.4**	**58.3**

achieved a high accuracy, combining our prompts learned with text data still brings further improvement on recognition performance. This also proves that texts and images are complementary to each other to some extent.

4.4. Integration with Partially Labeled Methods

Partially-labeled data refers to the problem where merely some labels of each sample are known. Following [5], we created the partial-label data by randomly masking out labels of the fully annotated data. During inference, the model is evaluated on all categories.

We reproduce DualCoOp [28] on partial-labeled VOC2007 and MS-COCO with the same experimental setting as reported (reproduced results are marked with *) and explore the enhancement brought by integration with our TaI-DPT. We also deploy DualCoOp on partial-labeled NUS-WIDE for a comprehensive comparison.

The results are reported in Table 3. With no prior knowledge from pre-trained models, previous forefront method like SARB [23] struggles to learn from incomplete labels. DualCoOp, directly trained with entire image set, is well-learned to classify multi-label images. Even so, for PascalVOC and NUS-WIDE, by using texts from exogenous sources, *i.e.*, COCO and OpenImages, our method can bring complementary enhancement to image prompting. While, for COCO, we used captions derived from its own images, making it more difficult to provide additional benefits. Analogous to VOC and NUS-WIDE, the COCO results can also be further advanced by introducing more external text data in future work.

4.5. Ablation Study

To thoroughly investigate the effect of each component, we conduct a series of ablation studies on the quantity of texts, training loss, different VL pre-trained models, ensemble weights, and texts *v.s.* images for prompting. The complete ablation experiments are shown in the *Suppl*.

Quantity of texts. Following the data preparation procedure in Sec. 3.2, we end up with a total number of 66087 pieces of text that contain descriptions for 20 categories of VOC2007. We test the performance of TaI-DPT with different numbers of randomly selected texts, and the results are shown in Fig. 7. When no collected texts are available, 80 templates of hand-crafted prompts from [24], like "a cropped photo of a [CLASS]", are used for training (all templates are shown in the *Suppl*), and each template sentence correlates with one positive label corresponding to the class name inserted in [CLASS]. The increasing number of texts gradually forms a complete description of target categories, and the relationship between classes is also better characterized, which results in ascending performance.

5. Conclusion

In this paper, we propose a new view of treating texts as images in prompt tuning (*i.e.* TaI), which learns the prompt from discriminative features of text descriptions. Compared to prior prompt tuning methods trained with images, our TaI benefits from the easy accessibility of scalable content-rich texts, which enables prompt tuning for vision tasks (*e.g.*, multi-label image recognition) even without downstream image data. Double-grained prompting is further introduced to utilize both the global and fine-grained features for better multi-label recognition ability. Nonetheless, when few-shot image samples or partial-labeled images are available, our TaI-DPT can conveniently integrate with existing prompting methods. Experiments on multiple benchmarks show the validity of TaI-DPT.

Acknowledgement

This work was supported in part by National Key R&D Program of China under Grant No. 2020AAA0104500, and by the National Natural Science Foundation of China (NSFC) under Grant No. U19A2073.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 1

[2] Amit Alfassy, Leonid Karlinsky, Amit Aides, Joseph Shtok, Sivan Harary, Rogerio Feris, Raja Giryes, and Alex M Bronstein. Laso: Label-set operations networks for multi-label few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6548–6557, 2019. 2, 7

[3] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification, 2021. 2

[4] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009. 4

[5] Tianshui Chen, Tao Pu, Hefeng Wu, Yuan Xie, and Liang Lin. Structured semantic transfer for multi-label recognition with partial labels. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 339–346, 2022. 2, 8

[6] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 522–531, 2019. 2, 6

[7] Zhao-Min Chen, Xiu-Shen Wei, Xin Jin, and Yanwen Guo. Multi-label image recognition with joint class-aware map disentangling and label correlation embedding. In *ICME 2019*. 2

[8] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *CVPR*, 2019. 2, 6

[9] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 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*, pages 1–9, 2009. 2, 6, 7

[10] 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. 2

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 6, 7

[12] Bin-Bin Gao and Hong-Yu Zhou. Multi-label image recognition with multi-class attentional regions. *arXiv preprint arXiv:2007.01755*, 2020. 2

[13] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *arXiv preprint arXiv:2202.06687*, 2022. 3

[14] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894*, 2013. 2, 6

[15] Shiyi He, Chang Xu, Tianyu Guo, Chao Xu, and Dacheng Tao. Reinforced multi-label image classification by exploring curriculum. In *AAAI*, 2018. 2

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1

[17] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022. 3

[18] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017. 2, 4, 6

[19] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 1

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 4, 6, 7

[21] Luchen Liu, Sheng Guo, Weilin Huang, and Matthew Scott. Decoupling category-wise independence and relevance with self-attention for multi-label image classification. In *ICASSP 2019*, 05. 2

[22] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5206–5215, 2022. 3

[23] 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. 2, 8

[24] 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. 1, 2, 4, 6, 7, 8

[25] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 82–91, 2021. 6

[26] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *arXiv preprint arXiv:2209.07511*, 2022. 3

[27] Christian Simon, Piotr Koniusz, and Mehrtash Harandi. Meta-learning for multi-label few-shot classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3951–3960, 2022. 2, 7

[28] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *arXiv preprint arXiv:2206.09541*, 2022. 1, 2, 3, 5, 6, 8

[29] Ashwin Vaswani, Gaurav Aggarwal, Praneeth Netrapalli, and Narayan G Hegde. All mistakes are not equal: Comprehensive hierarchy aware multi-label predictions (champ). *arXiv preprint arXiv:2206.08653*, 2022. 2

[30] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *CVPR*, 2016. 2

[31] Yangtao Wang, Yanzhao Xie, Yu Liu, Ke Zhou, and Xiaocui Li. Fast graph convolution network based multi-label image recognition via cross-modal fusion. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1575–1584, 2020. 2

[32] Yunchao Wei, Wei Xia, Min Lin, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. Hcp: A flexible cnn framework for multi-label image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38, 2015. 2

[33] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 1

[34] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021. 3

[35] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *ECCV*, 2020. 2

[36] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 1

[37] Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, and Jianfeng Lu. Multilabel image classification with regional latent semantic dependencies. *IEEE Transactions on Multimedia*, 20, 2018. 2

[38] Renrui Zhang, Zhang Wei, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. *arXiv preprint arXiv:2207.09519*, 2022. 7

[39] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Neural prompt search. *arXiv preprint arXiv:2206.04673*, 2022. 3

[40] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 1, 3

[41] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 3, 4, 6, 7

[42] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. *arXiv preprint arXiv:2205.14865*, 2022. 3