DualCoOp++: Fast and Effective Adaptation to Multi-Label Recognition with Limited Annotations

Ping Hu, Ximeng Sun, Stan Sclaroff, and Kate Saenko

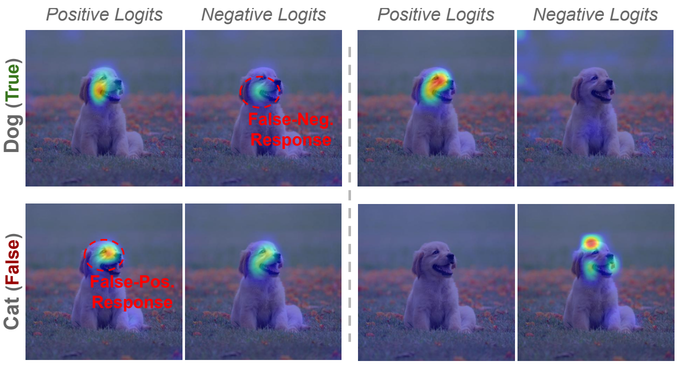
Abstract—Multi-label image recognition in the low-label regime is a task of great challenge and practical significance. Previous works have focused on learning the alignment between textual and visual spaces to compensate for limited image labels, yet may suffer from reduced accuracy due to the scarcity of high-quality multi-label annotations. In this research, we leverage the powerful alignment between textual and visual features pretrained with millions of auxiliary image-text pairs. We introduce an efficient and effective framework called *Evidence-guided Dual Context Optimization* (`DualCoOp++`), which serves as a unified approach for addressing partial-label and zero-shot multi-label recognition. In `DualCoOp++` we separately encode evidential, positive, and negative contexts for target classes as parametric components of the linguistic input (i.e., prompts). The evidential context aims to discover all the related visual content for the target class, and serves as guidance to aggregate positive and negative contexts from the spatial domain of the image, enabling better distinguishment between similar categories. Additionally, we introduce a Winner-Take-All module that promotes inter-class interaction during training, while avoiding the need for extra parameters and costs. As `DualCoOp++` imposes minimal additional learnable overhead on the pretrained vision-language framework, it enables rapid adaptation to multi-label recognition tasks with limited annotations and even unseen classes. Experiments on standard multi-label recognition benchmarks across two challenging low-label settings demonstrate the superior performance of our approach compared to state-of-the-art methods.

Index Terms—Multi-label image recognition, vision-language model, partial-label recognition, zero-shot recognition

1 Introduction

Image recognition has become a very popular and successful research area in recent years, due to the development of large-scale datasets [1], [2] and advanced model architectures [3], [4], [5], [6]. However, the majority of image recognition approaches have focused on single-label prediction, which ignores the intrinsic multi-label nature of images. Unlike single-label recognition [3], [4], [5], [6], multi-label image recognition aims to recognize all semantic labels present in an image [7], [8], [9], [10], [11], [12], [13], providing a more comprehensive understanding and benefiting applications like image retrieval, video analysis, and recommendation systems.

Multi-label recognition (MLR) typically deals with images of complex scenes and diverse objects. Collecting multi-label annotations becomes difficult to scale up, for two reasons: (i) annotating images with the full semantic label set is laborious and (ii) samples of particular categories can be hard to find. The first challenge can be addressed by multi-label recognition with *partial labels*, where merely some of the categories are annotated for each training image. Recent works proposed solutions to partial-label MLR based on semi-supervised learning [14], [15], normalized training objectives [16], or label correlations [17], [18], [19]. The second setting involves *zero-shot* MLR, where novel unseen categories are recognized by transferring knowledge from seen categories, with solutions like principal image features [20], [21], knowledge graphs [22], and attention mechanisms [23], [24]. Despite significant progress on the two settings, existing approaches are not designed to handle both at once. We propose to unify these settings as *limited-annotation* MLR and design a solution that can handle practical scenarios with either partial or missing labels.



(a) DualCoOp (b) DualCoOp++

Fig. 1: **Visualization of positive and negative logit maps for the true label "Dog" and the false label "Cat".** In contrast to `DualCoOp` showing high false-negative and false-positive activations, `DualCoOp++` presents better abilities to suppress incorrect predictions.

Successful solutions to the above problems transfer knowledge from fully-annotated categories to partially-labeled and novel categories by learning an alignment between images and category names [17], [19], [21]. Recently, vision-language pretraining models are bridging the visual-textual gap via large-scale pretraining, e.g., CLIP [26] is trained with 400 million image-text pairs. In this work, we

- *P. Hu, X-M. Sun, and S. Sclaroff are with the Department of Computer Science, Boston University, Boston, MA 02215.*
  E-mail: pinghu@bu.edu
- *K. Saenko are with Boston University, Boston, MA 02215 and Meta, Menlo Park, CA 94025.*

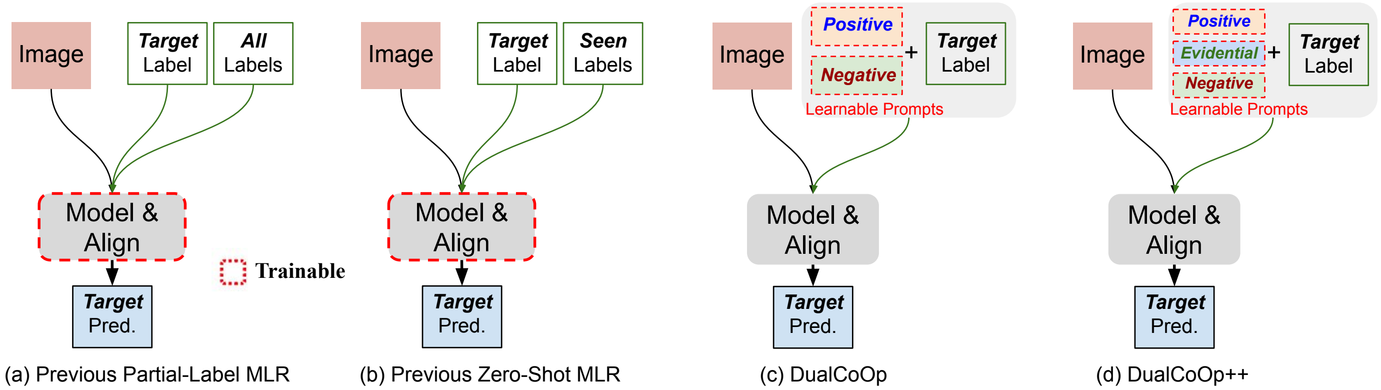


Fig. 2: **A conceptual comparison of previous multi-label recognition (MLR) methods and our approaches**. In Partial-Label MLR (a) and Zero-Shot MLR (b), previous works learn to model and align the visual and textual inputs as well as explore the correlation between the target label with all/seen labels depending on the limited semantic annotations available on the dataset, which leads to sub-optimal performance and complex model architectures. In contrast, we propose a unified framework (c) to tackle both limited-annotation tasks [25], and further improve the representation ability (d). We rely on the alignment of visual and textual inputs contained a large-scale pretrained vision-language model, and only learn an extra set of light-weight prompts to this model.

draw inspiration from the recent success of prompt learning for such models [27], [28], [29], [30], [31], [32], [33], [34], [35], [36]. Prompt learning provides a convenient way to transfer pretrained vision-language models to other tasks. It designs additional templated or learnable prompt tokens for textual input to "inform" the model about downstream tasks and avoids finetuning the entire model which can be inefficient and data-hungry. By doing so, recent works like CoOp [37] have demonstrated CLIP's remarkable generalization to various zero-shot image tasks [26], [27], [32]. However, these methods mainly focus on matching each image with a single label, hence they are not able to handle the multi-label setting.

To adapt the knowledge learned in CLIP to multi-label image recognition, we propose the `DualCoOp` in the conference version [25] of this work. As shown in Fig. 2 (c), `DualCoOp` learns a pair of differentiable prompts to provide positive and negative contexts for the target class. Instead of using hand-crafted thresholding to determine positive labels [38], the dual prompts naturally result in a positive and a negative classifier, so the existence of the target class in the image can be easily decided by comparing their scores. Unlike prior models, shown in Fig. 2 (a)(b), we avoid fine-tuning the full vision-language model and only learn the prompts, which are much smaller compared to the entire model. Therefore, the simple framework achieves much higher efficiency when adapting to different datasets.

In `DualCoOp` [25], we introduce the Class-Specific Region Feature Aggregation, where the positive context is directly normalized as the spatial attention to aggregate final positive and negative logits. Given samples with true labels (i.e. the image contains the target class), such a design can properly promote true-positive predictions by highlighting positive logits and suppressing negative logits. When learning with false labels (i.e. the image doesn't contain the target class), the objective aims to optimize true-negative predictions by minimizing positive logits and maximizing negative logits. However, minimizing the positive logit map suppresses the response in image regions of interest, which `DualCoOp` fully relies on for aggregation, resulting in distracted aggregation weights that make the learning of negative logits intractable. As a result, the learned model may still suffer from the loss of accuracy. To relieve this limitation, in this extended version of work, we propose `DualCoOp++` by introducing an Evidence-Guided Region Feature Aggregation module. As shown in Fig. 2(d), besides the positive and negative contexts, we further introduce the evidential context to guide the spatial aggregation of positive and negative contexts. Unlike positive logits and negative logits that directly indicate the existence and non-existence of object classes, evidential logits aim to extract all the related visual contents showing similar representations. As a result, optimizing the positive branch will not affect the learning of the negative branch, and the model can better represent and distinguish between target classes and similar classes as visualized in Fig. 1. Moreover, since `DualCoOp` is optimized for each class individually, the lack of interaction among classes may also hamper the performance. In particular, an image region can positively respond to multiple similar classes, hence resulting in false-positive predictions. To address this challenge, we further propose a Winner-Take-All (WTA) module that regularizes each spatial location to only positively respond to at most one category, thus further enhancing the model's ability to distinguish between similar categories. Since WTA is a non-parametric module, the proposed framework keeps high efficiency without introducing extra computational overhead. With these design choices, we achieve a unified framework for addressing the general challenges of multi-label recognition with limited annotations.

Our contributions can be summarised as follows:

- We propose `DualCoOp++` to efficiently and effectively adapt a powerful vision-language model to solve multi-label recognition tasks using limited annotations.
- We propose the Evidence-Guided Region Feature Aggregation module to improve the spatial aggregation of contextual information learnt from limited annotations.
- We propose the Winner-Take-All (WTA) module to promote the inter-class interaction in MLR, leading to better distinguish similar categories.
- We conduct extensive experiments and anlaysis on multiple benchmark datasets, and demonstrate that

DualCoOp++ achieves the state-of-the-art performance of both partial-label MLR and zero-shot MLR. Notably, without introducing extra computational overhead, DualCoOp++ consistently improves our previous DualCoOp by more than 2% for both tasks on benchmarks like MS-COCO and NUS-WIDE.

2 Related Works

Multi-Label Recognition with Limited Annotations. Multi-label image recognition has drawn increasing attention in past years. One straightforward solution to this problem is to individually learn a binary classifier for each category [39], [40], [41], which however does not consider correlations among labels. Hence, recent works have focused on incorporating semantic dependencies among labels via graph neural networks [7], [8], [13] or RNN/LSTM [9], [12], [42], [43]. Some work also considers the spatial distribution of labels in the image, and exploits object proposals [10], [44], [45] or attention mechanism [11], [43], [46] as a regularization to rectify the prediction. However, despite achieving significant progress, these methods require a large-scale and complete annotated dataset to train models [47], [48]. This limits their application to more practical scenarios where data is partially annotated for training [15], [49], [50], [51], [52], [53] and unseen (zero-shot) categories may appear during testing [21], [22], [54], [55], [56].

With partially labeled data, where merely some labels of each sample are known, Mahajan *et al.* [15] and Joulin *et al.* [14] attempt to use web supervision to automatically generate the pseudo labels, which unfortunately leads to poor performance as the web supervision is noisy and incomplete [57]. To avoid external noise, Durand *et al.* [16] exploit the proportion of annotated samples for different labels and propose a normalized BCE loss to train models based on the given partial labels. More recent works explicitly transfer information from known labels to complement unknown labels by utilizing category-specific feature blending [19] or label co-occurrences [17] at both instance-level and prototype-level.

Unlike partial annotation of the same label set for training and testing, zero-shot multi-label image recognition needs to handle novel categories during testing, hence inspiring a different route based on a joint visual-label embedding space [21], [23], [24], [54], [56]. Zhang *et al.* [21] propose to find a principal direction that ranks related labels first in the joint embedding space optimized via a tailored zero-shot ranking loss. Cohen *et al.* [20] further improve the idea by learning multiple principal vectors to support the semantic diversity. Huynh *et al.* [23] consider the spatial regularization and propose a shared multi-attention model and obviate the need for explicit region proposals [58]. Narayan *et al.* [24] propose to enhance the region-based features so as to minimize inter-class feature entanglement.

Though significant progress has been made in each of the directions, existing methods still require a lot of MLR data and complex architectures/losses. Our approach reduces the need for hard-to-get MLR data by pretraining on unsupervised text-image pairs. While it may seem unfair to compare existing MLR methods with ones based on such pretraining, we point out that the pretraining data is unsupervised and thus easier to obtain. We also provide experiments comparing DualCoOp++ to baselines using the same pretraining. Importantly, previous methods are designed for only one task, hence have limitations in practical applications. In contrast, our proposed framework can be easily adapted with small data and can address both partial and zero-shot tasks at the same time.

Prompt Learning for Vision-Language Models. Vision-Language Models [26], [59] based on contrastive learning have demonstrated an impressive ability to learn generic visual representations. As a milestone, CLIP [26] is trained with 400 million curated image-text pairs, and shows remarkable transfer capability for over 30 classification datasets. With such powerful vision-language models, several follow-ups [60], [61], [62], [63], [64], [65], [66] have been proposed to explore the training strategies for training downstream classification tasks. Instead of fine-tuning the entire model [4], [67], which may damage the learned representation space, recent approaches adopt the prompt-based paradigm that formalizes NLP tasks as masked language modeling (prompt templates) [68], [69], [70], [71], [72]. Zhou *et al.* [37] propose to tune prompts for downstream classification tasks, and further introduce input-conditional prompts for better generalization ability [32]. Lu *et al.* [73] learn the distribution of diverse prompts to handle the varying visual representations. Huang *et al.* [27] generate pseudo labels for images to learn prompts in an unsupervised way. Though achieving promising improvements for downstream tasks, these methods address the multi-class zero-shot image recognition, assuming each image has one label, hence lacking the ability to handle the multi-label setting, especially under the low-label regime. Toward this direction, Ding *et al.* [74] introduce a semantic correspondence prompt network to enhance the semantic context with label-to-label semantic priors, while Guo *et al.* [75] exploit rich text description data to learn stronger prompts. In contrast, without relying on extra networks or data, we propose a unified method that transfers vision-language models to address limited-annotation multi-label image recognition with better performance.

3 Method

Problem Definition. We formally define multi-label recognition with limited annotations as follows: Consider M as the set of categories which describe objects or attributes in images. Given a training image I, the existence of a category $m \in M$ can be positive, negative or unknown, corresponding to the label $y_m = 1, -1$ or 0 respectively. During inference, we predict each label of interest for an input image.

Many existing MLR problems fit into this broad definition. In this paper, we consider the settings with partial or missing labels: (1) **Partial-label MLR** [16], [17], [19], in which only a subset of labels are known ($+1$ or -1) for each training image and we are interested in predicting all existing labels during inference. (2) **Zero-shot MLR** [20], [23], [76], in which each label is either known (seen) or unknown (unseen) for *all* images during training and we are interested in predicting either all labels or only unknown (unseen) labels during inference. In this paper, we propose a unified setting that includes both scenarios, which we call *limited-annotation MLR*.

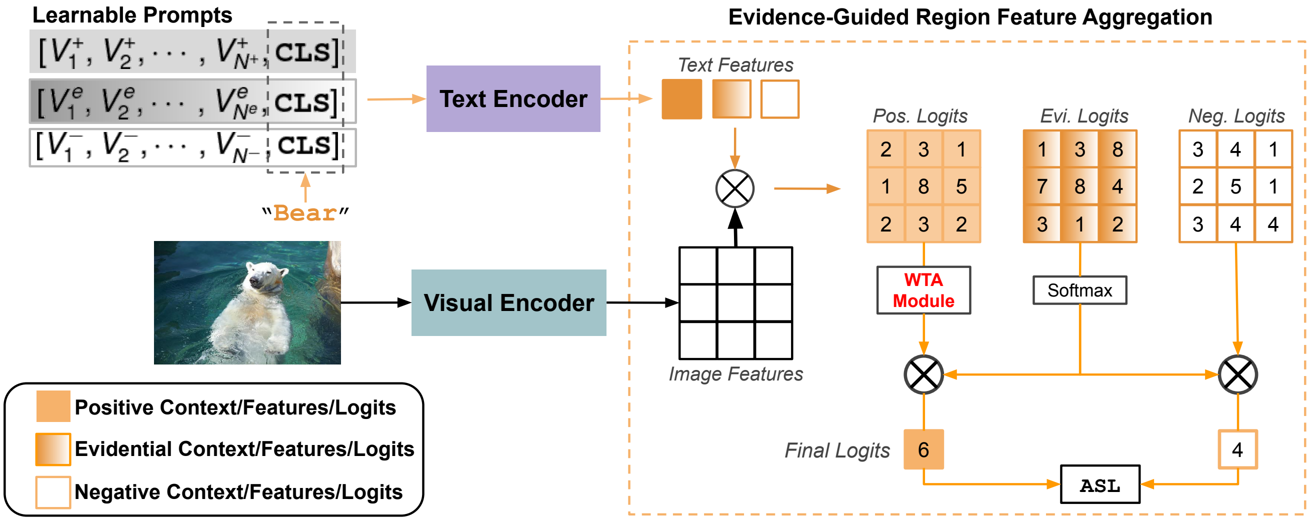

Learnable Prompts
$[V_1^+, V_2^+, \cdots, V_{N^+}^+,$ CLS]
$[V_1^e, V_2^e, \cdots, V_{N^e}^e,$ CLS]
$[V_1^-, V_2^-, \cdots, V_{N^-}^-,$ CLS]
"Bear"
Text Encoder
Visual Encoder
Evidence-Guided Region Feature Aggregation
Text Features
Image Features
Pos. Logits
Evi. Logits
Neg. Logits
2 3 1
1 8 5
2 3 2
1 3 8
7 8 4
3 1 2
3 4 1
2 5 1
3 4 4
WTA Module
Softmax
Final Logits
6
4
ASL
Positive Context/Features/Logits
Evidential Context/Features/Logits
Negative Context/Features/Logits

Fig. 3: **Illustration of our proposed approach.** `DualCoOp++` learns a triplet of evidential, positive, and negative prompts to quickly adapt powerful pretrained vision-text encoders to the MLR task. For each class, three prompts generate three contrastive (evidential, positive, and negative) textual embeddings as the input to the text encoder. Furthermore, we propose *Evidence-Guided Region Feature Aggregation* which projects each region's feature to the textual space first and then aggregate the spatial logits by the magnitude of class-specific evidential responses. An winner-take-all (WTA) module is also utilized to promote cross-class interaction. During training, we apply the ASL loss [38] to optimize learnable prompts while keeping other network components frozen. During inference, we compare the final positive and negative logits to make a prediction for each class.

Approach Overview. To compensate for insufficient or missing image labels, it is important to learn how the meanings of category names are related to each other, so that we can transfer knowledge between related categories. This is usually done by learning an alignment between the visual and textual spaces. However, our dataset is too limited to learn a broad and generalizable mapping. We propose to instead leverage the strong alignment of visual and textual feature spaces learned by large-scale vision-language pre-training (CLIP [30]) with a light-weight learnable overhead that quickly adapts to the MLR task with limited semantic annotations. Figure 3 provides an overview of our proposed approach. `DualCoOp++` learns a triplet of "prompt" contexts in the form of three learnable sequences of word vectors, to provide evidential, positive, and negative contexts of a given category name. This generates evidential, positive, and negative textual features that are fed into the pretrained text encoder. To better distinguish and recognize target objects, which can be located at different locations in the image and similar to other categories, we introduce an evidence-guided spatial aggregation step. We first compute the similarity score of the projected visual feature maps with the three context encodings to obtain three prediction logits over each region. For each class, we perform spatial aggregation of all positive/negative logits, in which the weight for each logit is determined by its relative magnitude of the evidential logits, and we call this *Evidence-Guided Region Feature Aggregation*. By doing so, the proposed framework can be more flexible to represent and distinguish the target class and similar classes (e.g. Regions of "Dog" can show high similarity to label "Cat" in the evidential logit map, while avoiding response in the positive logit map). We further apply a non-parametric Winner-Take-All module to promote inter-class interaction and optimize the learnable prompts via the ASL loss [38] while keeping all other network components frozen. During inference, we directly compare the final positive and negative logits to make a prediction for a category's label y.

Triple Learnable Prompts. Instead of learning a single [37] or dual [25] prompt(s) for a class, we propose Evidence-Guided Context Optimization (`DualCoOp++`) which learns three contrastive prompts' contexts for each class. The learnable part in triple prompts carries evidential, positive, and negative contextual surroundings individually and can be optimized end-to-end from data via binary classification loss. Specifically, we define the triplet of prompts given to the text encoder as follows:

$$P^e = [V_1^e,\ V_2^e, \cdots,\ V_{N^e}^e, \texttt{CLS}] \tag{1}$$

$$P^+ = [V_1^+, V_2^+, \cdots, V_{N^+}^+, \texttt{CLS}] \tag{2}$$

$$P^- = [V_1^-, V_2^-, \cdots, V_{N^-}^-, \texttt{CLS}] \tag{3}$$

where each V is a learnable word embedding vector (*e.g.* with dimension 512 in CLIP [30]) and `CLS` is the given category name. N^e, N^+, and N^- are the numbers of word tokens learned in the evidential, positive, and negative prompts respectively. For simplicity, we utilize the same size of prompts in our experiments. We learn a triplet of prompts for each class (i.e. class-specific prompt triplet) when solving MLR with partial labels, and learn a triplet of prompts shared for all classes in zero-shot MLR. With a triplet of prompts, we compute the binary classification output p by comparing the positive and negative contexts as:

$$p = \frac{e^{\delta(E_v^I, E_t^e, E_t^+)/\tau}}{e^{\delta(E_v^I, E_t^e, E_t^+)/\tau} + e^{\delta(E_v^I, E_t^e, E_t^-)/\tau}} \tag{4}$$

where p is the predicted probability for a given (image, label) pair as a positive example, and τ is a temperature parameter. E_v^I is the visual encoding feature maps. E_t^e, E_t^+, E_t^- are the textual encodings for evidential, positive, and negative prompts, respectively. $\delta(\cdot,\cdot,\cdot)$ is our proposed evidence-guided spatial aggregation function to adaptively reduce the spatial dimension of visual features for each class, which will be discussed next.

Evidence-Guided Region Feature Aggregation. In multi-label image recognition, it is common that multiple objects appear in different regions of the image. Pooling to produce a single image-level feature vector for all classes gives sub-optimal performance since spatial information is reduced and different objects are mixed. In this work, we reformulate the last multi-headed attention pooling layer of the visual encoder in CLIP [30] and apply evidence-guided class-specific pooling to adaptively aggregate region features in the multi-label setting. The original attention pooling layer in CLIP pools the visual feature map first, and then projects the global feature vector into text space as follows:

$$\begin{aligned}\text{AttnPool}(x) &= \text{Proj}_{v\to t}\Big(\sum_i \text{softmax}\big(\frac{q(\bar{x})k(x_i)^T}{C}\big)\cdot v(x_i)\Big)\\ &= \sum_i \text{softmax}\big(\frac{q(\bar{x})k(x_i)^T}{C}\big)\cdot \text{Proj}_{v\to t}(v(x_i))\\ &= \text{Pool}(\text{Proj}_{v\to t}(v(x_i))) \qquad (5)\end{aligned}$$

where q, v and k are independent linear embedding layers and $x = E_v^I$ is the output feature map of the visual encoder. By removing the pooling operation, we can project the visual feature F_v^i of each region i to the textual space [31]:

$$F_v^i = \text{Proj}_{v\to t}(v(E_i^I)) \qquad (6)$$

For each region i and a target class, we compute the logits with cosine similarity between F_v^i and the class's evidential, positive, and negative contexts,

$$S_i^e = \frac{F_v^i \cdot E_t^e}{||F_v^i||\cdot||E_t^e||} \qquad (7)$$

$$S_i^+ = \frac{F_v^i \cdot E_t^+}{||F_v^i||\cdot||E_t^+||} \qquad (8)$$

$$S_i^- = \frac{F_v^i \cdot E_t^-}{||F_v^i||\cdot||E_t^-||} \qquad (9)$$

In order to make a single prediction for the whole image, we aggregate the logit maps S_i^+ and S_i^- into S^+ and S^- according to the magnitude of S_i^e, and achieve the evidence-guided spatial aggregation:

$$\delta(E^I, E_t^e, E_t^+) = \sum_i \big(\text{softmax}(S_i^e)\cdot S_i^+\big) \qquad (10)$$

$$\delta(E^I, E_t^e, E_t^-) = \sum_i \big(\text{softmax}(S_i^e)\cdot S_i^-\big) \qquad (11)$$

Then, the prediction for the target class can be made by applying Eq. 4. Notably, we do not introduce any new parameters in our re-formulation of the spatial aggregation function. All parameters used to project visual features to the textual space are inherited from the original multi-headed attention pooling layer in CLIP.

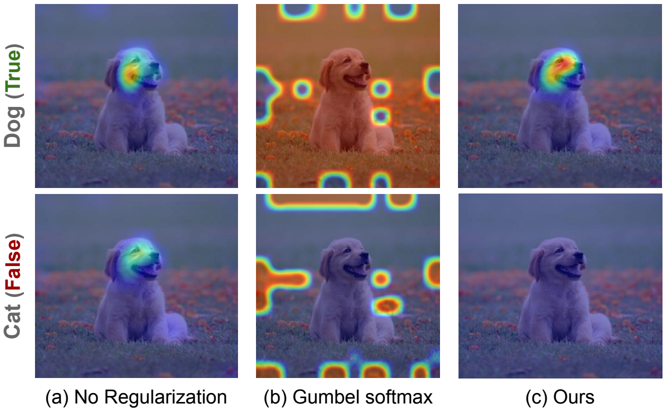


Fig. 4: **Positive logit maps for different types of inter-class regularization.** (a) Without any regularization, the head area of the dog shows large positive logits to both "Cat" and "Dog". (b) Gumbel softmax always highlights the larger logit, correlating class labels to background regions. (c) Our proposed WTA module can highlight true-positive logits and suppress false-positive logits.

Winner-Take-All Regularization. So far, the proposed framework learns to make predictions for different classes independently. Yet as shown in Fig. 4 (a), treating classes independently can result in false-positive predictions, as the same visual features may positively respond to multiple similar classes, especially under the limited-annotation regime where sufficient supervision is lacking. We address this issue with a Winner-Take-All (WTA) module that regularizes each spatial region to only positively respond to at most one class. Given an image region i and a label set containing M target categories, we have M positive logit scores $S_i^+ = [(S_i^+)^0, (S_i^+)^1, ..., (S_i^+)^{M-1}]$ based on Eq. 8. The regularization weights $w_i \in \mathcal{R}^M$ are computed over all the labels,

$$w_i = \text{softmax}\big(\gamma \cdot S_i^+ \cdot \max_m(S_i^+)\big) \qquad (12)$$

where $\max_m(S_i^+)$ represents the maximum logit score for regions i across the M classes, and γ is a hyperparameter.

Then, we update the positive logits elementwisely as,

$$(S_i^+)' = w_i \odot S_i^+ \qquad (13)$$

where $\odot$ is the Hadamard product. As illustrated in Fig. 4 (b)(c), in contrast to Gumbel softmax which always signalizes the maximum element, our WTA highlights the larger elements only if more than one logit has large values, which ensures that an image region can positively respond to none or just one of the given classes.

Optimization. We apply the Asymmetric Loss (ASL) [38] to handle the inherent positive-negative imbalance in the optimization of multi-label recognition. Specially, we compute losses for a positive (image, label) pair $\mathcal{L}_+$ and a negative (image, label) pair $\mathcal{L}_-$ as follows:

$$\mathcal{L}_+ = (1-p)^{\gamma_+}\log(p) \qquad (14)$$

$$\mathcal{L}_- = (p_c)^{\gamma_-}\log(1-p_c) \qquad (15)$$

where p is the probability in Eq. 4, and $p_c = \max(p-c, 0)$ is the probability for negative examples shifted by hard thresholding via the margin c. We set the hyper-parameters $\gamma_- \geq \gamma_+$, so that ASL down-weighs and hard-thresholds easy negative samples. The pair of learnable prompts are updated by back-propagating ASL through the frozen text encoder.

TABLE 1: **Multi-label Recognition on MS-COCO and VOC2007 with partial labels.** `DualCoOp++` achieves the best performance over all SOTA methods. * indicates previous models using weights pretrained by CLIP [26]

Methods	#P	10%	20%	30%	40%	50%	60%	70%	80%	90%	Avg.
MS-COCO [48]											
SSGRL [77]	64.7M	62.5	70.5	73.2	74.5	76.3	76.5	77.1	77.9	78.4	74.1
GCN-ML [7]	44.9M	63.8	70.9	72.8	74.0	76.7	77.1	77.3	78.3	78.6	74.4
KGGR [54]	≥ 25M	66.6	71.4	73.8	76.7	77.5	77.9	78.4	78.7	79.1	75.6
Curriculum labeling [16]	≥ 38M	26.7	31.8	51.5	65.4	70.0	71.9	74.0	77.4	78.0	60.7
Patial BCE [16]	≥ 38M	61.6	70.5	74.1	76.3	77.2	77.7	78.2	78.4	78.5	74.7
SST [17]	33.5M	68.1	73.5	75.9	77.3	78.1	78.9	79.2	79.6	79.9	76.7
SST*	33.5M	69.1	78.5	79.3	79.9	80.1	80.5	81.1	80.7	80.7	78.9
SARB [19]	29.6M	71.2	75.0	77.1	78.3	78.9	79.6	79.8	80.5	80.5	77.9
SARB*	29.6M	75.5	78.5	79.0	79.5	80.4	80.2	80.8	80.6	80.8	79.4
SCPNet [74]	3.4M	80.3	82.2	82.8	83.4	83.8	83.9	84.0	84.1	84.2	83.2
`DualCoOp` [25]	**1.3M**	78.7	80.9	81.7	82.0	82.5	82.7	82.8	83.0	83.1	81.9
`DualCoOp++`	1.5M	**81.4**	**83.1**	**83.7**	**84.2**	**84.4**	**84.5**	**84.8**	**85.0**	**85.1**	**84.0**
PASCAL VOC 2007 [78]											
SSGRL [77]	66.6M	77.7	87.6	89.9	90.7	91.4	91.8	91.9	92.2	92.2	89.5
GCN-ML [7]	44.9M	74.5	87.4	89.7	90.7	91.0	91.3	91.5	91.8	92.0	88.9
KGGR [54]	≥ 25M	81.3	88.1	89.9	90.4	91.2	91.3	91.5	91.6	91.8	89.7
Curriculum labeling [16]	≥ 38M	44.7	76.8	88.6	90.2	90.7	91.1	91.6	91.7	91.9	84.1
Patial BCE [16]	≥ 38M	80.7	88.4	89.9	90.7	91.2	91.8	92.3	92.4	92.5	90.0
SST [17]	32.4M	81.5	89.0	90.3	91.0	91.6	92.0	92.5	92.6	92.7	90.4
SARB [19]	29.6M	83.5	88.6	90.7	91.4	91.9	92.2	92.6	92.8	92.9	90.7
SPCNet [74]	–	91.1	92.8	93.5	93.6	93.8	94.0	94.1	94.2	94.3	93.5
`DualCoOp` [25]	**0.3M**	90.3	92.2	92.8	93.3	93.6	93.9	94.0	94.1	94.2	93.2
`DualCoOp++`	0.4M	**92.7**	**93.4**	**93.8**	**94.0**	**94.3**	**94.4**	**94.4**	**94.7**	**94.9**	**94.1**

4 Experiments

In this section, we first report the performance on partial-label and zero-shot multi-label recognition benchmarks, then present experiments to analyze the proposed method.

4.1 Multi-Label Recognition with Partial Labels

Datasets. We conduct experiments on MS-COCO [48], VOC2007 [78] and BigEarth [79] to evaluate multi-label recognition with partial labels. MS-COCO [48] contains 80 common object categories and we use the official `train2014` (82K images) and `val2014` (40K images) splits for training and test. VOC2007 [78] contains 20 object categories and we use the official `trainval` (5K images) and `test` (5K images) splits for training and test. Furthermore, since CLIP pretraining data is not publicly available and it is plausible that CLIP pretraining data covers many coarse and fine-grained visual domains since it performs well in the zero-shot evaluation for many downstream tasks, we also experiment on a Remote Sensing Image dataset BigEarth [79], whose domain is far from the domains of the datasets in the mainstream papers (i.e. PASCAL VOC, MS-COCO, and NUS-WIDE).

To create the training set with partial labels, we follow the standard practice [16], [17], [19] to mask out labels from the fully annotated training set, and use the remaining labels for training. The proportion of kept labels varies from 10% to 90% as in previous works [17], [19].

Evaluation. We report the mean average precision (mAP) for each proportion of labels available for optimization (from 10% to 90%) and its average value for all proportions. We count the learnable parameters (#P) of each baseline and `DualCoOp++` to measure the complexity of optimization.

Implementation. For fair comparison, we adopt ResNet-101 [4] as the visual encoder in all baselines and `DualCoOp++` with input resolution 448×448, and use the same Transformer [30], [80] in CLIP [26] as the text encoder. The visual and text encoders are initialized from the CLIP pretrained model and kept frozen during optimization. For each class/label, we learn three independent context vectors with 12 context tokens (N = 12) to keep a similar size of parameters to `DualCoOp`. Note that these context tokens are the only learnable parts in `DualCoOp++`. We use the SGD optimizer with an initial rate of 0.002 which is decayed by the cosine annealing rule. We train context vectors for 50 epochs with a batch-size 32/32/8 for MS-COCO/BigEarth/VOC2007, respectively. For ASL loss, we choose $\gamma_+ = 1$, $\gamma_- = 2$ and $c = 0.05$ via validation. Training is done with one RTX A6000.

Baselines. To evaluate the effectiveness of `DualCoOp++`, we compare with the following baselines:

- SSGRL [77], GCN-ML [7] and KGGR [54] that adopt graph neural networks for label dependencies.
- Curriculum labeling [16] and SST [17] that generate pseudo labels for unknown labels.
- Partial BCE [16] that uses a normalized BCE loss to better exploit partial labels.
- SARB [19] that blends category-specific representation across different images to transfer information.
- SCPNet [74], TaI-DPT [75], and our DualCoop [25] that adopt the large-scale pretrained vision-language model CLIP.

Results. Table 1 shows the comparison of mAP between `DualCoOp++` and all baselines optimized with 10% to 90% of labels. For the two recent works (SST [17] and SARB [19]), we further substitute the ImageNet pretrained weights [4] with the CLIP pretrained weights [26] when initializing of their visual encoders, which results in SST* and SARB* in Table 1. Since we learn class-specific prompts, `DualCoOp++` on

TABLE 2: **Comparison between TaI-DPT [75] and our methods on MS-COCO.** All follow the same training setting [75].

Method	#P	10%	30%	50%	70%	90%	Avg.
TaI-DPT	1.3M	81.5	83.3	83.9	84.2	84.5	83.5
DualCoOp	**1.3M**	81.0	82.9	83.5	84.0	84.3	83.1
DualCoOp++	1.5M	**81.9**	**84.1**	**84.6**	**85.0**	**85.4**	**84.2**

TABLE 3: **Comparison between SARB [19] and our methods on BigEarth.** All use parameters pretrained by CLIP [26].

Method	#P	10%	30%	50%	70%	90%	Avg.
SARB	29.6M	71.5	76.5	78.6	80.4	84.2	78.2
DualCoOp	**0.3M**	81.7	86.5	90.1	91.7	92.2	88.4
DualCoOp++	0.4M	**83.4**	**90.3**	**91.9**	**92.3**	**93.0**	**90.1**

MS-COCO adopts more learnable parameters than VOC2007. The proposed DualCoOp++ achieves the best performance across all proportions of labels available during the training. Notably, DualCoOp++ consistently improve DualCoOp with similar learnable overhead, and outperforms the second-best method SCPNet [74] on both MS-COCO and VOC2007 with less than half of the learnable parameters. Especially, when only providing 10% of labels during the training, DualCoOp++ improves DualCoOp by more than 2% and outperforms SPCNet by more than 1% on both datasets. We also adopt the training protocols in TaI-DPT [75], which uses a larger batchsize and more tunable hyperparameters, and compare the results in Table. 2. Without extra training data, DualCoOp++ can further boost the performance and consistently outperform TaI-DPT, which exploits rich captioning data for training and two CLIP models for inference. This indicates DualCoOp++'s ability to quickly adapt to the multi-label recognition task with a few labels. On BigEarth, we compare DualCoOp++ with DualCoOp and a strong baseline SARB. Table 3 shows that DualCoOp++ consistently improves over DualCoOp and SARB with significant gaps under different portions of labels, which proves DualCoOp++ boosts the performance in various visual domains by taking advantage of the powerful vision-language pretraining.

Full Label Training. On MS-COCO, we also learn with 100% of training labels. Without finetuning the visual encoder, DualCoOp++ achieves 85.3% mAP, which improves DualCoOp by 2% and outperforming previous SOTA approaches like ASL [38] (85.0% mAP) and CSRA [84] (83.5% mAP) with the same ResNet-101 backbone. This shows DualCoOp++'s promising ability to exploit the pretrained CLIP model for addressing the challenging MLR tasks.

Computational Cost. We compare the computational cost between DualCoOp++ and previous methods in terms of training/testing latency and memory (see Table 4) using the same device (one Nividia A100 GPU). For the current multi-label recognition task, the categories are pre-set before inference, (i.e. we already know which class we would like to consider during inference.) In this case, we compute the text features for each class from the learned prompts and the class name ahead of the inference. Then we use the pre-computed text features to predict each image during the test. Since the text features are pre-computed (very light-weighted computing overhead), the text encoder is not executed during inference. For inference, the latency time and memory consumption of DualCoOp++ are the same as DualCoOp when using the same backbone for the image encoder. During training, CLIP-based methods slightly raise latency time and memory consumption since image and text encoders are both executed during the forward, and only prompts are updated in DualCoOp++.

TABLE 4: **Computations Efficiency Comparison.**

Methods	Training		Testing	
	Latency ms/img	Memory GB/img	Latency ms/img	Memory GB/img
SARB [19]	4.7	0.21	4.0	0.13
TaI-DPT [75]	–	–	4.8	0.09
DualCoOp [25]	5.3	0.22	4.0	0.06
DualCoOp++	6.6	0.22	4.0	0.06

4.2 Zero-shot Multi-Label Recognition

Datasets. Following [20], [23], we conduct experiments on MS-COCO [48] and NUS-WIDE [8] to perform zero-shot multi-label recognition. On MS-COCO, we follow [20], [85] to split the dataset into 48 seen classes and 17 unseen classes. NUS-WIDE [8] dataset includes 270K images. Following [20], [23] we use 81 human-annotated categories as unseen classes and an additional set of 925 labels obtained from Flickr tags as seen classes.

Evaluation. We follow [20] and report precision, recall, and F1 score at Top-3 predictions in each image on MS-COCO. We also follow [20], [23] to report mAP over all categories as well as precision, recall, and F1 score at Top-3 and Top-5 predictions in each image on NUS-WIDE. We evaluate all methods with both zero-shot setting (test only on unseen classes) and generalized zero-shot setting (test on both seen and unseen classes).

Implementation. We adopt ResNet-50 [4] similar to [20] as the visual encoder in DualCoOp++ for input resolution 224. Instead of learning class-specific prompts, we learn the class-agnostic context vectors with 42 context tokens (N = 42) for all classes, which is the only learnable part in DualCoOp++. We optimize context vectors for 50 epochs with a batch-size 32/192 for MS-COCO/NUS-WIDE, respectively. During inference, we combine the learnt pair of context vectors with the class name for each class (either base class or novel class) and compute the text features. Other implementation details are the same as in Sec. 4.1

Baselines. To evaluate the effectiveness of DualCoOp++ in the zero-shot setting, we compare with the following baselines:

- CONSE [81] that adopts an ensemble of classifiers for unseen classes.
- LabelEM [82] that learns a joint image-label embedding.
- Fast0Tag [21] and SDL [20] that estimate one or multiple diverse principal directions of the input images.
- Deep0Tag [76] and LESA [23] that estimate the relevant regions via region proposals and attention techniques respectively.
- BiAM [24] that enhances the region-based features to minimize inter-class feature entanglement.
- DualCoOp [25] that is based on the pretrained CLIP model.

TABLE 5: **Zero-Shot Multi-label Recognition on NUS-WIDE.** `DualCoOp++` achieves the best F1 score over all SOTA methods at Top-3/Top-5 predictions in both ZSL and GZSL settings.

Methods	#P	Zero-Shot Learning (ZSL)							Zero-Shot Learning (GZSL)						
		Top-3			Top-5			mAP	Top-3			Top-5			mAP
		P	**R**	**F1**	**P**	**R**	**F1**		**P**	**R**	**F1**	**P**	**R**	**F1**	
CONSE [81]	-	17.5	28.0	21.6	13.9	37.0	20.2	9.4	11.5	5.1	7.0	9.6	7.1	8.1	2.1
LabelEM [82]	-	15.6	25.0	19.2	13.4	35.7	19.5	7.1	15.5	6.8	9.5	13.4	9.8	11.3	2.2
Fast0Tag [21]	0.61M	22.6	36.2	27.8	18.2	48.4	26.4	15.1	18.8	8.3	11.5	15.9	11.7	13.5	3.7
OAL [83]	$\geq$ 12.8M	20.9	33.5	25.8	16.2	43.2	23.6	10.4	17.9	7.9	10.9	15.6	11.5	13.2	3.7
LESA$_{M10}$ [23]	$\geq$0.45M	25.7	41.1	31.6	19.7	52.5	28.7	19.4	23.6	10.4	14.4	19.8	14.6	16.8	5.6
BiAM [24]	3.8M	–	–	33.1	–	–	30.7	26.3	–	–	16.1	–	–	19.0	9.3
SDL$_{M7}$ [20]	33.6M	24.2	41.3	30.5	18.8	53.4	27.8	25.9	27.7	13.9	18.5	23.0	19.3	21.0	12.1
`DualCoOp` [25]	**0.07M**	37.3	46.2	41.3	28.7	59.3	38.7	43.6	31.9	13.9	19.4	26.2	19.1	22.1	12.0
`DualCoOp++`	**0.07M**	**42.4**	**52.5**	**46.9**	**31.2**	**64.5**	**42.1**	**47.1**	**34.7**	**15.2**	**21.1**	**29.2**	**21.3**	**24.7**	**15.1**

TABLE 6: **Zero-Shot Multi-Label Recognition on MS-COCO**. `DualCoOp++` achieves the best F1 score in both ZSL and GZSL settings.

Methods	ZSL			GZSL		
	P	**R**	**F1**	**P**	**R**	**F1**
CONSE [81]	11.4	28.3	16.2	23.8	28.8	26.1
Fast0Tag [21]	24.7	61.4	25.3	38.5	46.5	42.1
Deep0Tag [76]	26.5	65.9	37.8	43.2	52.2	47.3
SDL$_{M2}$ [20]	26.3	65.3	37.5	59.0	60.8	59.9
`DualCoOp` [25]	35.3	87.6	50.3	58.4	68.1	62.9
`DualCoOp++`	**36.8**	**91.4**	**52.5**	**59.4**	**69.3**	**64.0**

TABLE 8: **Partial-label MLR performance with 50% annotations.** "Evi." represents the evidence-guided spatial aggregation. "WTA" denotes the winner-take-all module in training.

Dataset	Evi.	WTA	C_P	C_R	C_F	O_P	O_R	O_F
MSCOCO			72.1	80.4	75.8	73.7	83.9	78.5
	✓		74.3	80.3	77.0	76.1	83.7	79.7
	✓	✓	**76.0**	**80.9**	**77.9**	**77.0**	**84.1**	**80.4**
VOC			80.6	93.4	86.3	82.4	94.0	87.8
	✓		81.8	93.1	86.4	83.1	93.9	88.2
	✓	✓	**82.6**	**93.5**	**87.5**	**84.2**	**94.1**	**88.9**

TABLE 9: **Zero-shot MLR performance.** "Evi." represents the evidence-guided spatial aggregation. "WTA" denotes the winner-take-all module in training.

			ZSL			GZSL		
	Evi.	WTA	**P**	**R**	**F1**	**P**	**R**	**F1**
MS-COCO			35.3	87.6	50.3	58.4	68.1	62.9
	✓		35.9	89.2	51.2	59.0	68.9	63.6
	✓	✓	**36.8**	**91.4**	**52.5**	**59.4**	**69.3**	**64.0**
NUS-WIDE			37.3	46.2	41.3	31.9	13.9	19.4
	✓		37.9	46.3	41.7	32.8	14.3	19.9
	✓	✓	**42.4**	**52.5**	**46.9**	**34.7**	**15.2**	**21.1**

Results. Table 5 and 6 show the comparison between `DualCoOp++` and all SOTA methods of zero-shot learning and generalized zero-shot learning on NUS-WIDE and MS-COCO datasets. `DualCoOp++` achieves the best accuracy in all cases with a very light learnable overhead (0.07M) and improves the performance of zero-shot learning with significant margins. Compared to previous state-of-the-art SDL [20], `DualCoOp++` improves ZSL performance by 15.0 @Top-3 on MS-COCO, and by 14.1 @Top-3 and 14.3 @Top-5 on NUS-WIDE. This shows the power of exploiting the pretrained alignment of textual and visual spaces in CLIP via `DualCoOp++` to solve multi-label recognition. `DualCoOp++` also consistently improves the precision and recall of `DualCoOp` in all the settings, demonstrating the effectiveness of our proposed methods to suppress false predictions.

4.3 Method Analysis

Effectiveness of Text Supervision. To show the effectiveness of text supervision from label space, we compare the model learned with discrete label space ("Discrete Label") with four methods (SST [17], SARB [19], CoOp [37] and `DualCoOp++`) which introduce the textual space to utilize the contextual correlation of labels in Table 7. We find that methods with text supervision usually perform better than the method only using discrete labels. However, when the semantic annotations are limited, text supervision sometimes yields worse performance (e.g. mAP of SST is 1.5% lower than Discrete Labels with only 10% of labels). CoOp [37] utilizes the visual-textual alignment. However, with the original multi-head attention and single positive prompt, it yields worse performance than Discrete Labels. To better utilize the well-pretrained alignment for MLR tasks, `DualCoOp++` learns a context triplet and adopts evidence-guided region feature aggregation, which leads to great performance (*e.g.* 10.8% higher than Discrete Labels with 10% of labels) and quickly adapts to the dataset even with limited labels.

TABLE 7: **Comparison among methods on MS-COCO using partial labels with the same initialization. All methods use parameters pretrained by CLIP [26].**

Methods	Text Sup.	10%	30%	50%	70%	90%
Disc. Label	✗	70.6	75.1	76.5	77.3	78.0
SST	✓	69.1	79.3	80.1	81.1	80.7
SARB	✓	75.5	79.0	80.4	80.8	80.8
CoOp	✓	63.0	68.5	69.2	71.5	75.0
`DualCoOp`	✓	78.7	81.7	82.5	82.8	83.1
`DualCoOp++`	✓	**81.4**	**83.7**	**84.4**	**84.8**	**85.1**

`DualCoOp++` v.s. `DualCoOp`. We analyze the impact of the newly introduced components in `DualCoOp++` in Table. 8 and 9. For partial-label recognition, we report the per-class and the average overall precision (C_P and O_P), recall (C_R and O_R), and F1 measure (C_F and O_F). As shown in Table 8, introducing the evidence-guided spatial aggregation (denoted as "Evi.") improves the precision significantly and

TABLE 10: **Ablation on Linguistic Inputs for Zero-Shot Learning of MS-COCO.** M2∼M5 have similar sizes of learnable parameters.

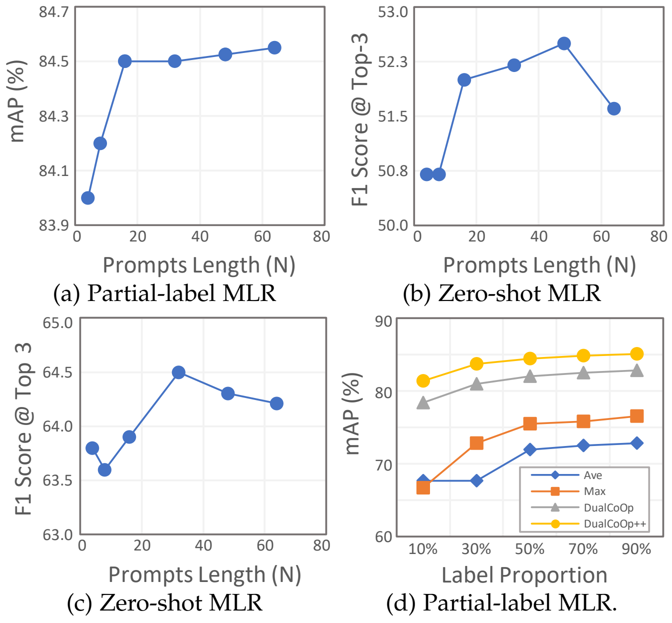
	Linguistic Input	ZSL			GZSL		
		P	**R**	**F1**	**P**	**R**	**F1**
M0	Contextless	5.2	12.9	7.4	3.5	4.1	3.8
M1	Hand-crafted	25.6	63.6	36.5	31.0	36.2	33.4
M2	Neg. Prompt	9.5	23.1	13.5	2.5	3.3	2.8
M3	Pos. Prompt	32.1	77.8	45.4	56.2	65.3	60.4
M4	Pos.+Neg. Prompts	35.3	87.6	50.3	58.4	68.1	62.9
M5	Evi.+Pos.+Neg. Prompts	**35.9**	**89.2**	**51.2**	**59.0**	**68.9**	**63.6**

TABLE 11: **Comparison between multi-headed attention and class-specific feature aggregation on MS-COCO**

Visual Aggre.	Fine-tune.	Train - Test Resolution	10%	30%	50%	70%	90%
Multi Headed Atten. [80]	✗	224 – 224	70.4	74.1	74.8	75.4	75.7
	✗	224 – 448	65.9	70.2	71.2	72.0	72.1
	✗	448 – 448	72.1	75.5	76.5	77.1	77.3
	✓	448 – 448	74.1	77.6	78.2	78.5	78.4
Evi.-Guided Spatial Aggre.	✗	224 – 224	76.2	78.7	79.3	79.7	80.2
	✗	224 – 448	78.2	79.5	81.3	82.1	83.0
	✗	448 – 448	81.4	83.7	84.4	84.8	85.1
	✓	448 – 448	81.9	84.1	85.2	85.2	85.7

F-measure, and further applying the winner-take-all module can boost the performance even higher. This demonstrates that our method can effectively relieve the limitation of `DualCoOp` and suppress false positive predictions. For zero-shot recognition as shown in Tab. 9, our proposed components show consistent improvements for all the metrics and datasets, demonstrating the ability to generalize to unseen classes.

Ablation of Prompt Design. We compare our proposed triple learnable prompts with hand-crafted prompts and single prompt learning method on the MS-COCO dataset with the zero-shot setting (see Table 10). Hand-crafted prompts can use either contextless class names [86] or manually designed prompt templates. In our experiments, we carefully choose the positive and negative prompt templates as "a photo of a [classname]" and "a photo without a [classname]". In contrast with performing the binary classification for each class with multiple learnable prompts as the input, we also perform experiments with learning a single prompt of positive or negative contexts and use a chosen threshold (0.5 in our experiment) to make the prediction for each class. As we can see, the single positive prompt learning method (M3) performs better than non-learnable methods (M0 and M1), and a single negative learnable prompt (M2) achieves much worse accuracy than its positive counterpart (M3). Including both positive and negative prompts (M4) performs better than a single prompt. When adding the evidential prompt, triple prompts (M5) achieves even higher accuracy, which indicates that `DualCoOp++` can better handle the complementary and beneficial information from positive and negative sides. In Fig. 5(a)(b)(c), we analyze the impact of different lengths of prompt context in all three different experiment scenarios. In MLR with partial labels, we learn class-specific prompts and thus a smaller N (e.g. 12) in `DualCoOp++` can perform well. For zero-shot learning in MLR, we learn uniform prompts shared by all classes and it requires a larger N (e.g. 36) for good performance. In the main paper, we use N = 12 for all experiments of MLR with partial labels and use N = 42 for experiments in zero-shot learning to keep a similar size to `DualCoOp`.



(a) Partial-label MLR

(b) Zero-shot MLR

(c) Zero-shot MLR

(d) Partial-label MLR.

Fig. 5: Analysis of different design options on MS-COCO.

Impact of Aggregation Function. In Table 11, we compare the adaptive ability of different spatial aggregation methods. In general, training/testing with a larger resolution is beneficial as spatial details matter. As shown, multi-headed attention is bonded to the pre-training image resolution (224 in CLIP), while our evidence-guided region aggregation benefits from the increased input resolution either during training or in inference. Our method uses original weights, but actually performs much better than finetuning the original multi-headed attention layer. We also compare different functions to aggregate the regional logits for each class in Fig. 5 (d). We compute final logits in four ways: (1) taking the average of logits at all spatial locations ("Ave"), (2) taking the region with the largest positive logit ("Max"), (3) generating aggregating weights for all spatial locations via a softmax function over the positive logits ("DualCoOp"), (4) generating aggregating weights for all spatial locations via a softmax function over the evidential logits ("DualCoOp++"). "Max" performs better than "Ave", which indicates the regional feature is more informative than the global feature in multi-label recognition. By taking account of both the regional and the global features, "DualCoOp" improves performance. When further introducing the evidential, our "DualCoOp++" gives the best performance, which demonstrates that generating aggregating weights over the evidential logits is better than aggregating via positive logits. Visualization of different logits in `DualCoOp` and `DualCoOp++` is presented in Fig. 6.

Ablation of Finetuning DualCoOp. To better examine the effectiveness of finetuning for `DualCoOp++`, we also conduct experiments to finetune the weights in evidence-guided region aggregation. As shown in Table 11, finetuning aggregation function can stably improve the performance over

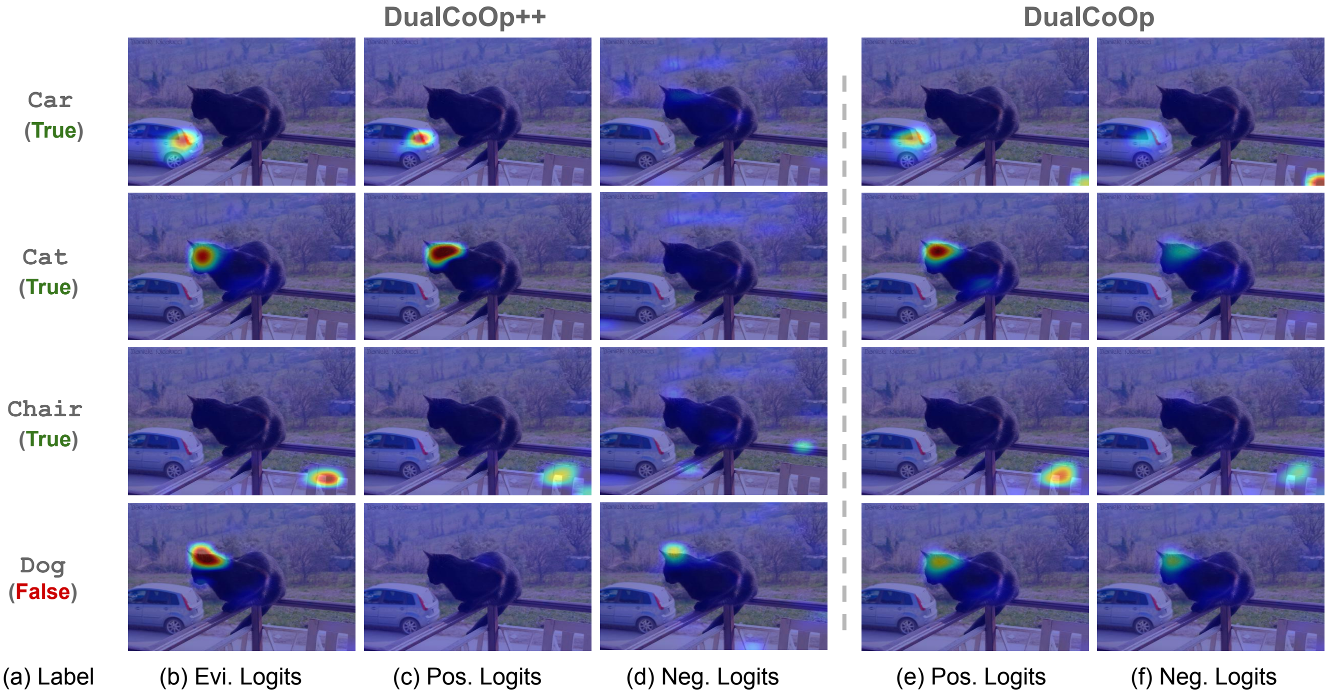


Fig. 6: **Visualization of logit maps in DualCoOp++ and DualCoOp**. Given class labels (a) to DualCoOp++, the evidential logit map (b) highlights closely related image regions, and the positive (c)/ negative (d) logit maps provided correct positive and negative support for the highlighted areas to make final predictions. While in DualCoOp, negative samples are not well optimized hence leading to weak negative logit maps, resulting in false-positive prediction (i.e. higher positive response than the negative response in the cat region for the label Dog).

the non-finetuning setting with different amounts of labels available. We also tried finetuning all weights in the CLIP image encoder, yet find that performance drops significantly especially when given a lower portion of training labels, which shows that tuning with insufficient supervision may undermine the pretrained vision-language alignment in CLIP models.

5 Conclusion

In this paper, we extend our previous DualCoOp with a novel framework, DualCoOp++, unified for two types of multi-label recognition with limited annotations: partial-label and zero-shot. DualCoOp++ utilizes powerful vision-language pretraining model obtained from a web-scale dataset. By introducing a lightweight learnable overhead, it can quickly adapt to solve multi-label recognition after receiving a small amount of labels. In DualCoOp++, we learn a triplet of evidential, positive, and negative prompts followed by the target class name as the linguistic input. Furthermore, to better aggregate visual region features for each class, we reformulate the original visual attention in the pretraining model as an evidence-guided region feature aggregation. Moreover, a winner-take-all module is introduced to promote the cross-label interaction and regularize that each region positively responds to at most one class. We conduct extensive experiments for both partial-label MLR and Zero-Shot MLR across MS-COCO, VOC2007, and NUS-WIDE datasets, showing the improvements of DualCoOp++ to DualCoOp and the efficacy of our proposed approach over state-of-the-art methods.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*. Ieee, 2009.

[2] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov *et al.*, "The open images dataset v4," *IJCV*, vol. 128, no. 7, pp. 1956–1981, 2020.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[5] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[7] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *CVPR*, 2019.

[8] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "Nus-wide: a real-world web image database from national university of singapore," in *ICIVR*, 2009.

[9] F. Liu, T. Xiang, T. M. Hospedales, W. Yang, and C. Sun, "Semantic regularisation for recurrent image annotation," in *CVPR*, 2017.

[10] Y. Liu, L. Sheng, J. Shao, J. Yan, S. Xiang, and C. Pan, "Multi-label image classification via knowledge distillation from weakly-supervised detection," in *ACM MM*, 2018.

[11] N. Sarafianos, X. Xu, and I. A. Kakadiaris, "Deep imbalanced attribute classification using visual attention aggregation," in *ECCV*, 2018.

[12] V. O. Yazici, A. Gonzalez-Garcia, A. Ramisa, B. Twardowski, and J. v. d. Weijer, "Orderless recurrent models for multi-label classification," in *CVPR*, 2020.

[13] Y. Wang, D. He, F. Li, X. Long, Z. Zhou, J. Ma, and S. Wen, "Multi-label classification with label graph superimposing," in *AAAI*, 2020.

[14] A. Joulin, L. v. d. Maaten, A. Jabri, and N. Vasilache, "Learning visual features from large weakly supervised data," in *ECCV*, 2016.

[15] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. Van Der Maaten, "Exploring the limits of weakly supervised pretraining," in *ECCV*, 2018.

[16] T. Durand, N. Mehrasa, and G. Mori, "Learning a deep convnet for multi-label classification with partial labels," in *CVPR*, 2019.

[17] T. Chen, T. Pu, H. Wu, Y. Xie, and L. Lin, "Structured semantic transfer for multi-label recognition with partial labels," 2022.

[18] D. Huynh and E. Elhamifar, "Interactive multi-label cnn learning with partial labels," in *CVPR*, 2020.

[19] T. Pu, T. Chen, H. Wu, and L. Lin, "Semantic-aware representation blending for multi-label image recognition with partial labels," in *AAAI*, 2022.

[20] A. Ben-Cohen, N. Zamir, E. Ben-Baruch, I. Friedman, and L. Zelnik-Manor, "Semantic diversity learning for zero-shot multi-label classification," in *ICCV*, 2021.

[21] Y. Zhang, B. Gong, and M. Shah, "Fast zero-shot image tagging," in *CVPR*, 2016.

[22] C.-W. Lee, W. Fang, C.-K. Yeh, and Y.-C. F. Wang, "Multi-label zero-shot learning with structured knowledge graphs," in *CVPR*, 2018.

[23] D. Huynh and E. Elhamifar, "A shared multi-attention framework for multi-label zero-shot learning," in *CVPR*, 2020.

[24] S. Narayan, A. Gupta, S. Khan, F. S. Khan, L. Shao, and M. Shah, "Discriminative region-based multi-label zero-shot learning," in *ICCV*, 2021.

[25] X. Sun, P. Hu, and K. Saenko, "Dualcoop: Fast adaptation to multi-label recognition with limited annotations," *NeurIPS*, 2022.

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[27] T. Huang, J. Chu, and F. Wei, "Unsupervised prompt learning for vision-language models," *arXiv preprint arXiv:2204.03649*, 2022.

[28] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie, "Prompting visual-language models for efficient video understanding," *arXiv preprint arXiv:2112.04478*, 2021.

[29] T. Lüddecke and A. S. Ecker, "Prompt-based multi-modal image segmentation," *arXiv preprint arXiv:2112.10003*, 2021.

[30] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, 2019.

[31] C. Zhou, C. C. Loy, and B. Dai, "Denseclip: Extract free dense labels from clip," *arXiv preprint arXiv:2112.01071*, 2021.

[32] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," *arXiv preprint arXiv:2203.05557*, 2022.

[33] Y. Zang, W. Li, K. Zhou, C. Huang, and C. C. Loy, "Unified vision and language prompt learning," *arXiv preprint arXiv:2210.07225*, 2022.

[34] C. Ma, Y. Liu, J. Deng, L. Xie, W. Dong, and C. Xu, "Understanding and mitigating overfitting in prompt tuning for vision-language models," *IEEE Trans. on CSVT*, 2023.

[35] Y. Liu, Y. Lu, H. Liu, Y. An, Z. Xu, Z. Yao, B. Zhang, Z. Xiong, and C. Gui, "Hierarchical prompt learning for multi-task learning," in *CVPR*, 2023.

[36] S. T. Wasim, M. Naseer, S. Khan, F. S. Khan, and M. Shah, "Vita-clip: Video and text adaptive clip via multimodal prompting," in *CVPR*, 2023.

[37] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *arXiv preprint arXiv:2109.01134*, 2021.

[38] T. Ridnik, E. Ben-Baruch, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor, "Asymmetric loss for multi-label classification," in *ICCV*, 2021.

[39] W. Liu and I. Tsang, "On the optimality of classifier chain for multi-label classification," *NIPS*, vol. 28, 2015.

[40] I. Misra, C. Lawrence Zitnick, M. Mitchell, and R. Girshick, "Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels," in *CVPR*, 2016.

[41] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *International Journal of Data Warehousing and Mining*, vol. 3, no. 3, pp. 1–13, 2007.

[42] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "Cnn-rnn: A unified framework for multi-label image classification," in *CVPR*, 2016.

[43] Z. Wang, T. Chen, G. Li, R. Xu, and L. Lin, "Multi-label image recognition by recurrently discovering attentional regions," in *ICCV*, 2017.

[44] Y. Li, C. Huang, C. C. Loy, and X. Tang, "Human attribute recognition by deep hierarchical contexts," in *ECCV*, 2016.

[45] M. Wang, C. Luo, R. Hong, J. Tang, and J. Feng, "Beyond object proposals: Random crop pooling for multi-label image recognition," *IEEE Trans on IP*, vol. 25, no. 12, pp. 5678–5688, 2016.

[46] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in *CVPR*, 2017.

[47] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, vol. 123, no. 1, pp. 32–73, 2017.

[48] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[49] S. S. Bucak, R. Jin, and A. K. Jain, "Multi-label learning with incomplete class assignments," in *CVPR*, 2011.

[50] M. Chen, A. Zheng, and K. Weinberger, "Fast image tagging," in *ICML*, 2013.

[51] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *ICCV*, 2017.

[52] Y.-Y. Sun, Y. Zhang, and Z.-H. Zhou, "Multi-label learning with weak label," in *AAAI*, 2010.

[53] M.-K. Xie and S.-J. Huang, "Partial multi-label learning," in *AAAI*, 2018.

[54] T. Chen, L. Lin, X. Hui, R. Chen, and H. Wu, "Knowledge-guided multi-label few-shot learning for general image recognition," *IEEE Trans. on PAMI*, 2020.

[55] A. Gupta, S. Narayan, S. Khan, F. S. Khan, L. Shao, and J. van de Weijer, "Generative multi-label zero-shot learning," *arXiv preprint arXiv:2101.11606*, 2021.

[56] T. Mensink, E. Gavves, and C. G. Snoek, "Costa: Co-occurrence statistics for zero-shot classification," in *CVPR*, 2014.

[57] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[58] S. Rahman, S. Khan, and N. Barnes, "Deep0tag: Deep multiple instance learning for zero-shot image tagging," *IEEE Trans. on MM*, vol. 22, no. 1, pp. 242–255, 2019.

[59] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, 2021.

[60] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *arXiv preprint arXiv:2110.04544*, 2021.

[61] M. Wortsman, G. Ilharco, M. Li, J. W. Kim, H. Hajishirzi, A. Farhadi, H. Namkoong, and L. Schmidt, "Robust fine-tuning of zero-shot models," *arXiv preprint arXiv:2109.01903*, 2021.

[62] Y. Yao, A. Zhang, Z. Zhang, Z. Liu, T.-S. Chua, and M. Sun, "Cpt: Colorful prompt tuning for pre-trained vision-language models," *arXiv preprint arXiv:2109.11797*, 2021.

[63] R. Zhang, R. Fang, P. Gao, W. Zhang, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-adapter: Training-free clip-adapter for better vision-language modeling," *arXiv preprint arXiv:2111.03930*, 2021.

[64] M. M. Derakhshani, E. Sanchez, A. Bulat, V. G. T. da Costa, C. G. Snoek, G. Tzimiropoulos, and B. Martinez, "Variational prompt tuning improves generalization of vision-language models," *arXiv preprint arXiv:2210.02390*, 2022.

[65] X. Liu, D. Wang, M. Li, Z. Duan, Y. Xu, B. Chen, and M. Zhou, "Patch-token aligned bayesian prompt learning for vision-language models," *arXiv preprint arXiv:2303.09100*, 2023.

[66] A. Miyai, Q. Yu, G. Irie, and K. Aizawa, "Locoop: Few-shot out-of-distribution detection via prompt learning," *arXiv preprint arXiv:2306.01293*, 2023.

[67] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H.-W. Hon, "Unified language model pre-training for natural language understanding and generation," *NeurIPS*, 2019.

[68] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," *arXiv preprint arXiv:2104.08691*, 2021.

[69] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," *arXiv preprint arXiv:2101.00190*, 2021.

[70] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh, "Autoprompt: Eliciting knowledge from language models with automatically generated prompts," *arXiv preprint arXiv:2010.15980*, 2020.

[71] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *arXiv preprint arXiv:2304.00685*, 2023.

[72] B. X. Yu, J. Chang, H. Wang, L. Liu, S. Wang, Z. Wang, J. Lin, L. Xie, H. Li, Z. Lin *et al.*, "Visual tuning," *arXiv preprint arXiv:2305.06061*, 2023.

[73] Y. Lu, J. Liu, Y. Zhang, Y. Liu, and X. Tian, "Prompt distribution learning," *arXiv preprint arXiv:2205.03340*, 2022.

[74] Z. Ding, A. Wang, H. Chen, Q. Zhang, P. Liu, Y. Bao, W. Yan, and J. Han, "Exploring structured semantic prior for multi label recognition with incomplete labels," in *CVPR*, 2023.

[75] Z. Guo, B. Dong, Z. Ji, J. Bai, Y. Guo, and W. Zuo, "Texts as images in prompt tuning for multi-label image recognition," in *CVPR*, 2023.

[76] S. Rahman and S. Khan, "Deep multiple instance learning for zero-shot image tagging," in *ACCV*, 2018.

[77] T. Chen, M. Xu, X. Hui, H. Wu, and L. Lin, "Learning semantic-specific graph representation for multi-label image recognition," in *ICCV*, 2019.

[78] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.

[79] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 1144–1158, 2018.

[80] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NIPS*, 2017.

[81] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean, "Zero-shot learning by convex combination of semantic embeddings," *arXiv preprint arXiv:1312.5650*, 2013.

[82] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for image classification," *IEEE Trans. on PAMI*, vol. 38, no. 7, pp. 1425–1438, 2015.

[83] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear Attention Networks," in *NIPS*, 2018.

[84] K. Zhu and J. Wu, "Residual attention: A simple but effective method for multi-label recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 184–193.

[85] A. Bansal, K. Sikka, G. Sharma, R. Chellappa, and A. Divakaran, "Zero-shot object detection," in *ECCV*, 2018.

[86] A. Li, A. Jabri, A. Joulin, and L. Van Der Maaten, "Learning visual n-grams from web data," in *ICCV*, 2017.