Language-Driven Multi-Label Zero-Shot Learning with Semantic Granularity

Shouwen Wang[1,2], Qian Wan[3], Junbin Gao[4], Zhigang Zeng[1,2*]
[1] School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[2] Key Laboratory of Image Processing and Intelligent Control, Ministry of Education
[3] Medical University of Vienna [4] National University of Singapore
wangshouwen533@gmail.com, w252086746@gmail.com, gao.junbin.cn@gmail.com, zgzeng@hust.edu.cn

Abstract

Recent methods learn class-unified prompt contexts by image data to adapt CLIP to zero-shot multi-label image classification, which achieves impressive performance. However, simply tuning prompts is insufficient to deal with novel classes across different semantic granularity levels. This limitation arises due to the sparse semantic detail in prompt class names and the hierarchical granularity competition among class names caused by CLIP's contrastive loss. We propose a language-driven zero-shot multi-label learning framework to bridge associations among classes across multiple granularity levels through class name reconstruction. To achieve this, we first leverage a language model to generate structured text descriptions for each class, which explicitly capture (1) visual attributes, (2) hierarchical relationships, and (3) co-occurrence scenes. With the enriched descriptions, we then learn class names by extracting and aligning semantic relationships and features from them in the CLIP's shared image-text embedding space. Furthermore, we consider that similar text descriptions among different classes may introduce ambiguities. We mitigate these ambiguities by imposing a pair-based loss on learnable class names to enhance their distinctiveness. During inference, we aggregate semantic predictions from multiple image snippets to reinforce the identification of classes across different granularity levels. Comprehensive experiments demonstrate that our method surpasses state-of-the-art methods in multi-label zero-shot learning and effectively handles novel classes across different granularity levels.

1. Introduction

Many Multi-label classification (MLC) works [8, 10, 20, 38, 40, 41, 43, 46] have achieved remarkable classification performance, but their model training requires a large number of images with high-quality annotations. To reduce reliance on annotation data, multi-label zero-shot learning (ML-ZSL) is proposed to learn the transfer mechanism from seen classes to unseen (novel) classes.

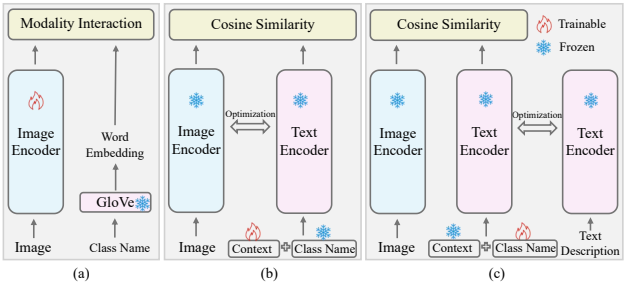

Figure 1. Comparison of classic multi-label zero-shot learning frameworks and our language-driven method. (a) represents the traditional separate alignment, while (b) and (c) illustrate the current joint alignment. (a) and (b) require annotated image data for training, whereas our method relies solely on text descriptions.

Most ML-ZSL studies achieve transferability by aligning visual and textual spaces. Pioneer works [3, 17, 21, 29, 45] utilize separate images and text features for alignment (Fig. 1(a)), with word embeddings of both seen and unseen classes generated by a pre-trained language-based model, *e.g.*, GloVe[32]. These methods depend solely on image data of seen classes to fine-tune the image encoder for aligning the separate textual space, significantly limiting transferability. Large-scale vision-language pre-trained models, *e.g.,* CLIP [33], are trained by image-text pairs, which has shown impressive transferability. Recent ML-ZSL studies [15, 16, 37] utilize CLIP's joint image-text alignment (Fig. 1(b)) to learn class-unified prompt contexts, improving cross-category generalizability, even on large-scale datasets such as NUS-WIDE [11] and Open Images [18].

Indeed, classes in large-scale multi-label datasets exhibit varying levels of semantic granularity, since the web-sourced images in these datasets are freely annotated by users with their own interests. For any given image, one user might focus on fine-grained classes (*e.g.*, "Golden Re-

* Corresponding author.

triever"), while another may emphasize broader, coarse-grained categories (*e.g.*, "dog" or even "animal"). Ideally, predictions should remain consistent across different granularity levels. However, without considering the semantic richness of class names, tuning prompts in prior ML-ZSL methods fail to establish associations among classes at multiple granularity levels, making it challenging to handle the semantic hierarchy in MLC. There are two main reasons: 1) In the "context + class name" paradigm, relying solely on the class name results in a lack of additional semantic information. This is especially problematic for class names that are coarsely defined (e.g., "organism", "urban") or abstractly defined (e.g., "travel", "happiness"), as they tend to be uninformative. 2) The contrastive loss of the pre-trained CLIP causes class names with hierarchical granularity (*e.g.*, "dog" and "animal") to compete when matching the same visual features, potentially leading to semantic inconsistencies. Therefore, our focus is on enriching the semantics of class names to handle the multi-label semantic granularity while maintaining the zero-shot transfer capability of CLIP.

To this end, we propose a language-driven zero-shot multi-label learning framework, which is illustrated in Fig. 1(c). Unlike previous zero-shot learning frameworks trained on image data, our framework only requires text data. Specifically, we first design prompts based on visual attributes, hierarchical relationships, and co-occurrence scenes of a class, and then query the language model GPT-4o mini [5] for text descriptions. These class-related descriptions explicitly highlight visual specificity and establish inherent hierarchical and contextual relationships. Next, we extract semantic relationships and features from the generated text descriptions to learn class names. However, relying solely on text for training can create a modality gap between textual and visual representations. To bridge the modality gap, the distance between text descriptions and learnable class names is optimized within the shared image-text embedding space. Moreover, similar text descriptions among different classes may introduce ambiguities, making it difficult to distinguish between them. To improve discrimination, we impose a pair-based loss between learnable class names. Reconstructed class names enriched with semantic knowledge are encoded to enhance the recognition of diverse class granularities. To further facilitate the matching of visual features with class embeddings at various granularity levels, we propose multi-snippet semantic aggregation across scales, enhancing the complementarity between global and local semantic prediction.

Our main contributions are summarized as follows:

1. We propose a novel framework of language-driven multi-label zero-shot learning with semantic granularity, which is only trained on text data. To the best of our knowledge, this is the first work to focus on zero-shot multi-label image classification across various semantic granularities.
2. The method can obtain **R**econstructed **C**lass **N**ames with rich sema**n**tics (**RCNn**) by class-related text descriptions, increasing the interpretability of the method. The proposed pair-based loss makes class names more distinguishable. Furthermore, we introduce a multi-snippet semantic aggregation module to improve the recognition of classes at different granularity levels.
3. Extensive experimental results demonstrate that our method significantly outperforms state-of-the-art approaches in the zero-shot learning task. For the generalized zero-shot learning task, we treat seen classes as novel classes, and our method still outperforms most approaches trained on image data. We also demonstrate the effectiveness of our approach in handling semantic granularity problems. *https://github.com/wangshouwen/RCNn*

2. Related work

Multi-Label Zero-Shot Learning. Traditional multi-label classification methods [8, 10, 20, 38, 40, 41, 43, 46] rely on large amounts of labeled image data. To reduce this dependency, alternative approaches such as multi-label partial-label learning [9, 39] and multi-label zero-shot learning (ML-ZSL) [17, 37] have been proposed. Recent ML-ZSL methods can be broadly categorized into vision-driven and language-driven approaches.

Vision-driven ML-ZSL typically utilizes the image supervision of seen classes to align the image-text space, enabling zero-shot transfer capabilities. Fast0Tag [45] and SDL [3] learn per-image principal directions to align the directions along which relevant labels rank ahead of irrelevant labels in the word vector space. LESA [17] generates multiple shared and label-agnostic attention maps for the selection of each label semantic vector. Considering that such shared maps lead to diffused attention, BiAM [29] enriches region-level features and maps them to class semantics for each class. Structured knowledge graphs [21] and ML-Decoder [35] use GNN and cross-attention for modality interaction, respectively. Despite their differences, these methods share a common limitation: they depend on a separate alignment approach, where textual embeddings are independently encoded by a pre-trained language model. The emergence of CLIP introduces a more integrated solution by enabling joint alignment of visual and textual spaces. Leveraging this capability, recent ML-ZSL methods such as DualCoOp [37] and MKT [16] optimize prompts to enhance image-text matching performance. However, despite this advancement, these approaches still require labeled images of seen classes for training. This reliance becomes particularly challenging for large-scale datasets (*e.g.,* NUS-WIDE [11], OpenImages [18]), where the extensive number of seen classes results in prohibitively high annotation costs.

To reduce such costs, language-driven ML-ZSL methods are proposed. TaI-DPT [15] utilizes text data instead

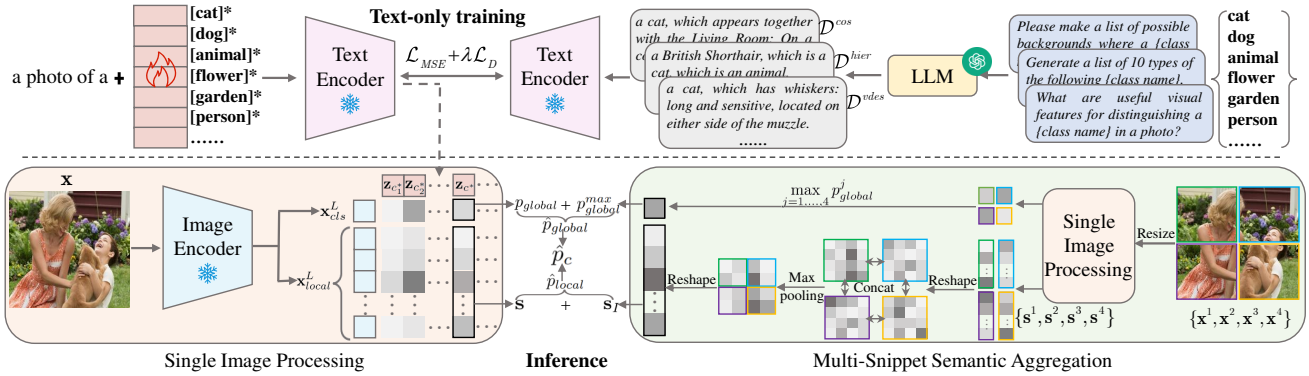

Figure 2. Illustration of our language-driven framework. Our backbone comprises pre-trained CLIP image and text encoders, which remain frozen. During training (Top), we reconstruct class names using semantic knowledge extracted from class-related text descriptions, which are generated by LLMs. During inference (Bottom), we align the global and local tokens of image features with the class embedding to generate predictions for class c. To better enable the recognition of classes across different granularity levels, we aggregate semantic predictions from multiple snippets through both global and local perspectives, with only the aggregation of class c shown.

of images for prompt tuning. Similarly, CoMC [26] trains a cross-modal classifier using language data. These methods use captions that describe image content and their derived class labels for supervised learning. In contrast, our method utilizes class-related textual descriptions to reconstruct class names, enhancing their semantic richness. We further consider the semantic granularity relationships existing within categories.

Semantic Hierarchy. Semantic hierarchy, based on a tree-like taxonomy [42] or a DAG-like semantic concept structure [36], has been widely utilized in prior studies to enhance various vision tasks [2, 4, 12, 13]. Building on this, leveraging label hierarchies has emerged as a promising strategy for addressing the zero-shot learning challenge, with applications extending to zero-shot multi-label text classification [6, 24]. In addition, hierarchical semantic-visual adaptation training [7, 44] effectively tackles separate feature alignment issues in zero-shot learning, whereas recent approaches such as CHiLS [31] and H-CLIP [14] harness semantic hierarchy to improve CLIP's zero-shot performance. Different from previous methods, our approach removes these constraints, which neither requires hierarchical annotations nor a specific semantic hierarchy, by leveraging hierarchical descriptions generated by large language models (LLMs) [5] to enrich class names dynamically.

3. Method

3.1. Problem Definition

For the traditional ML-ZSL problem [16, 29], it is necessary to define a seen class set and an unseen class set. The seen class set $\mathcal{C}^S$ includes categories annotated during training, whereas the unseen class set $\mathcal{C}^U$ consists of categories absent from the training annotations. The language-driven ML-ZSL aims to improve the class embeddings used as a classifier through text data without training with image data. let $\mathcal{D}=\{(\mathbf{t}_i,c_i)\}_{i=1}^M$ denote a training text corpus, where $\mathbf{t}_i$ is a text-based training sample, $c_i\in\mathcal{C}$ is its corresponding category, and M is the total number of text samples. $\mathcal{C}$ is the vocabulary of novel classes. For the language-driven ML-ZSL, both $\mathcal{C}^S$ and $\mathcal{C}^U$ are novel class sets. The evaluation is conducted on the standard zero-shot learning (ZSL) and the generalized zero-shot learning (GZSL) tasks. The scope of category for ZSL is $\mathcal{C}^U$, *i.e.*, $\mathcal{C}=\mathcal{C}^U$, and for GZSL is both $\mathcal{C}^S$ and $\mathcal{C}^U$, *i.e.*, $\mathcal{C}=\mathcal{C}^S\cup\mathcal{C}^U$.

3.2. Preliminary

CLIP model [33] includes an image encoder and a text encoder, and both encoders are trained on a large number of image-text pairs to align with each other. For the ViT-based image encoder with L residual attention blocks, the forward propagation of the L-th block is formulated as

$$\begin{aligned}
&\boldsymbol{q}=F_q(LN(\mathbf{x}_{L-1})),\boldsymbol{k}=F_k(LN(\mathbf{x}_{L-1})),\\
&\boldsymbol{v}=F_v(LN(\mathbf{x}_{L-1})),Attn_{qk}=Softmax(\frac{\boldsymbol{q}\boldsymbol{k}^\top}{\sqrt{d_k}})\\
&\bar{\mathbf{x}}=\mathbf{x}_{L-1}+\mathbf{x}_{attn}=\mathbf{x}_{L-1}+F(Attn_{qk}\cdot\boldsymbol{v}),\\
&\mathbf{x}_L=\bar{\mathbf{x}}+FFN(LN(\bar{\mathbf{x}})),
\end{aligned}\tag{1}$$

where $\mathbf{x}_{L-1}$ is the output of the $(L-1)$-th block, F denotes a linear layer, LN represents the layer normalization, and FFN is a feed-forward network. $Attn_{qk}$ means the self-attention of *query* and *key*, where d_k stands for the dimension of $\boldsymbol{k}$. $\mathbf{x}_L=[\mathbf{x}^L_{cls},\mathbf{x}^L_{patch}]$ is the output of the ViT architecture, where the class token $\mathbf{x}^L_{cls}\in\mathbb{R}^{1\times d}$ is used as the global feature of the image to align with the text embedding, and $\mathbf{x}^L_{patch}\in\mathbb{R}^{N\times d}$ represents the N local patch tokens in the d-dimensional space.

The alignment of the class token in CLIP is not friendly to the MLC task of identifying multiple categories. The class token tends to be dominated by salient categories, leading to the neglect of discriminative local features. ClearCLIP [19] tries to align local tokens with the text embedding. It makes three modifications to the L-th block: cutting the residual connection, selecting the self-attention of *query* and *query*, and discarding FFN. After modification, the matrix $\mathbf{x}_{local}^{L} \in \mathbb{R}^{N \times d}$ of the local tokens aligned with the text embedding is denoted as

$$\begin{aligned} \mathbf{x}_{attn} &= [\mathbf{x}_{cls}^{attn}, \mathbf{x}_{patch}^{attn}] = F(Attn_{qq} \cdot \boldsymbol{v}), \\ \mathbf{x}_{local}^{L} &= \mathbf{x}_{patch}^{attn}. \end{aligned} \tag{2}$$

For a candidate class c, we use the class token $\mathbf{x}_{cls}^{L}$ and the local tokens $\mathbf{x}_{local}^{L}$ to predict its logit score on the image $\mathbf{x}$ in the inference stage. The CLIP's text encoder E_t encodes the context prompt $t(\cdot)$ (*e.g.*, "`a photo of a` $\{\cdot\}$") for c and outputs the normalized class embedding $\mathbf{z}_c \in \mathbb{R}^{1 \times d}$, *i.e.*, $\mathbf{z}_c = E_t(t(c))$. The logit score p_{global} for the class token is as follows:

$$\begin{aligned} \tilde{\mathbf{x}}_{cls}^{L} &= Norm(F_p(LN(\mathbf{x}_{cls}^{L}))), \\ p_{global} &= \tilde{\mathbf{x}}_{cls}^{L} \cdot \mathbf{z}_c^{\top}, \end{aligned} \tag{3}$$

where F_p projects visual features into the image-text shared space and $Norm$ represents the $L2$-norm. The aggregated logit score p_{local} for the local tokens is as follows:

$$\begin{aligned} \tilde{\mathbf{x}}_{local}^{L} &= Norm(F_p(LN(\mathbf{x}_{local}^{L}))), \\ \boldsymbol{s} = \tilde{\mathbf{x}}_{local}^{L} \cdot \mathbf{z}_c^{\top}, p_{local} &= \sum_{i=1}^{N} softmax(s_i/\tau_s) s_i, \end{aligned} \tag{4}$$

where the score map $\boldsymbol{s} = [s_1, s_2, \ldots, s_N]^{\top}$ represents the logit scores of the local tokens for class c and τ_s is is a temperature coefficient. Finally, the logit score p_c of class c is computed as

$$p_c = (p_{local} + p_{global})/2. \tag{5}$$

3.3. Text Description Generation Based on LLM

A large-scale dataset (*e.g.*, NUS-WIDE, Open Image) contains hundreds of categories with rich semantic granularity. Such a large number of categories and different levels of semantic granularity have to be considered for annotating each image, which is an incredible labor. The language-driven ML-ZSL eliminates the dependence on annotated images through text training. Moreover, the emergence of large language models (LLMs), such as ChatGPT [5], makes it easier to collect text data. **It is worth considering what text data should be collected to cope with rich semantic granularity**.

Class names often provide limited information, particularly for coarse-grained, abstract, or scene-related classes. Relying solely on class names to identify novel classes with rich semantic granularity is difficult, and also inconsistent with human cognition. We humans generally start from visual features, semantic hierarchies, and co-occurrence scenes to summarize key features to determine categories. Inspired by this, we collect text data based on the above three aspects to reconstruct class names. We query a large language model (*e.g.*, GPT-4o mini) through the prompts containing class names to generate text descriptions as training samples. In particular, this generation method is not restricted by category. Our training text corpus $\mathcal{D}$ consists of class-related text descriptions, which encapsulate visual attributes, hierarchical relationships, and co-occurrence scenes. Details of text description generation are provided in Section B of the supplementary material.

3.4. Class Name Reconstruction

A class name alone containing limited information makes it difficult to cope with the semantic granularity problem of MLC; it is necessary to use the class-related knowledge from the descriptions to enrich the semantics of class names. A straightforward approach [25, 28] is to directly integrate the embeddings of all class-related text descriptions into a unified embedding for each class, such as mean and principal eigenvector. However, the mean does not consider the importance of significant features, and the principal eigenvector overlooks many details. We choose the method of learning class names to obtain a better embedding for each class, as shown in the training phase of Fig. 2. To maintain alignment with visual features, the class names are reconstructed in the shared image-text embedding space.

In the "context + class name" paradigm of CLIP, the prompt for class c can be denoted as $t(c) = \{v_1, v_2, \ldots, v_{N_t}, v_c\}$, where $\{v_1, v_2, \ldots, v_{N_t}\}$ are the context prompt vectors and N_t is the number of context vectors. $v_c = \{v_c^1, v_c^2, \ldots, v_c^{N_c}\}$ is the corresponding embedding set of the class name tokens and N_c is the token length, which is just a word-mapped embedding set of the class name and contains limited information. We aim to reconstruct class names using semantic descriptions of visual features, semantic hierarchies, and co-occurrence scenes. The learnable class name vectors $v_{c^*} = \{v_{c^*}^1, v_{c^*}^2, \ldots, v_{c^*}^{N_{c^*}}\}$ are introduced to learn the optimal class name c^*, where c^* and class c correspond one to one, and N_{c^*} is the same hyperparameter for all classes. The context prompt "`a photo of a`" and the learnable class name c^* are integrated to obtain the prompt $t(c^*) = \{v_1, v_2, \ldots, v_{N_t}, v_{c^*}\}$, where $N_t = 4$. The collected class-related text descriptions are used to learn v_{c^*} for each class. For a set $\mathcal{D}_{\mathcal{B}} \subseteq \mathcal{D}$ of the training text corpus, a text sample $\mathbf{t}_i$ and its corresponding class prompt $t(c_i^*)$ are projected into the shared image-text embedding space by CLIP's text encoder $E_t(\cdot)$. We extract semantic knowledge from the text description into the learn-

able class name by minimizing the Euclidean distance between the two embeddings in the shared space, as follows:

$$\mathcal{L}_{MSE} = \sum_{(\mathbf{t}_i, c_i) \in \mathcal{D}_{\mathcal{B}}} ||E_t(t(c_i^*)) - E_t(\mathbf{t}_i)||_2^2. \quad (6)$$

Generally, it is assumed that the discriminative information carried by class names serves as the foundation for identifying each class. Although knowledge extraction from text descriptions improves the semantic richness of class names, it cannot guarantee the distinguishability between classes. For similar classes, especially fine-grained classes (*e.g.*, Husky and Shiba), their text descriptions from three perspectives are very similar, so the learned class names contain limited discriminative information. To enhance the discrimination between classes, we introduce a pair-based loss for regularization in the shared space, formulated as

$$\mathcal{L}_D = \sum_{c \in \mathcal{C}} \sum_{\hat{c} \in \mathcal{C}} (1 + \frac{E_t(t(c^*)) \cdot E_t(t(\hat{c}^*))}{||E_t(t(c^*))||_2 \cdot ||E_t(t(\hat{c}^*))||_2}). \quad (7)$$

Thus, the overall object optimized by the collected text corpus is a combination of the above two losses, that is,

$$\mathcal{L} = \mathcal{L}_{MSE} + \lambda \mathcal{L}_D, \quad (8)$$

where λ is a balance hyperparameter. v_{c^*} for each class is optimized by minimizing $\mathcal{L}$.

3.5. Multi-Snippet Semantic Aggregation

In the inference phase of Fig. 2, the optimized class names are encoded into class embeddings to identify multiple labels for each image. For the MLC task with semantic granularity, coarse-grained and scene classes focus more on the global feature of an image, whereas fine-grained and object classes focus more on the discriminative local features of an image. Therefore, we integrate the predictions from both the class token and local tokens of an image to generate the final prediction for each class in Eq. 5. To enhance the complementarity of global and local image features, we further propose multi-snippet semantic aggregation (MSSA) in the scale dimension. MSSA improves the alignment of class embeddings and enhances the adaptability to various semantic granularities.

Specifically, MSSA is a split-to-integrate strategy. An image $\mathbf{x}$ is cropped into K snippets, denoted as $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^K\}$. Each snippet is resized back to the size of $\mathbf{x}$ and fed into CLIP's image encoder. The cosine similarities between the class tokens and local tokens of these snippets and the class embedding $\mathbf{z}_{c^*} = E_t(t(c^*))$ for class c are $\{p^1_{global}, p^2_{global}, \ldots, p^K_{global}\}$ and $\{\mathbf{s}^1, \mathbf{s}^2, \ldots, \mathbf{s}^K\}$ respectively, which are calculated by Eq. 3 and Eq. 4. Each $\mathbf{s}^j \in \mathbb{R}^{N \times 1}$ is reshaped into a semantic map with height H and width W, *i.e.*, $N = H \times W$. Leveraging spatial invariance, the semantic map of each snippet is concatenated into $\mathbf{s}_I$ according to its spatial position in the original image, formulated as $\mathbf{s}_I = Concat(\{\mathbf{s}^1, \mathbf{s}^2, \ldots, \mathbf{s}^K\})$ and $\mathbf{s}_I \in \mathbb{R}^{\sqrt{K}H \times \sqrt{K}W}$. To enhance the discriminability of local tokens, we perform semantic fusion at different scales. First, we ensure that $\mathbf{s}_I$ and $\mathbf{s}$ share the same dimensions by applying max pooling with a kernel size of $(\sqrt{K}, \sqrt{K})$ on $\mathbf{s}_I$. The pooled $\mathbf{s}_I$ is then reshaped to $N \times 1$. Finally, we aggregate the semantic information from both $\mathbf{s}_I$ and $\mathbf{s}$, denoted as $\hat{\mathbf{s}} = (\mathbf{s}_I + \mathbf{s})/2$. After aggregating the semantics of multi-snippet class tokens and local tokens, the logit scores $\hat{p}_{global}$ and $\hat{p}_{local}$ are computed as

$$\begin{aligned} \hat{p}_{global} &= (\max_{j=1,\ldots,K} p^j_{global} + p_{global})/2, \\ \hat{p}_{local} &= \sum_{i=1}^{N} softmax(\hat{s}_i/\tau_s)\hat{s}_i. \end{aligned} \quad (9)$$

Thus, Eq. 5 is further rewritten as

$$\hat{p}_c = (\hat{p}_{global} + \hat{p}_{local})/2. \quad (10)$$

Methods	ZSL			GZSL		
	P	**R**	**F1**	**P**	**R**	**F1**
CONSE [30]	11.4	28.3	16.2	23.8	28.8	26.1
Fast0Tag [45]	24.7	61.4	25.3	38.5	46.5	42.1
Deep0Tag [34]	26.5	65.9	37.8	43.2	52.2	47.3
SDL (M=2) [3]	26.3	65.3	37.5	**59.0**	60.8	59.9
DualCoOp [37]	<u>35.3</u>	<u>87.6</u>	<u>50.3</u>	<u>58.4</u>	**68.1**	**62.9**
TagCLIP* [23]	33.0	81.9	47.0	45.1	52.7	48.6
RCNn(ours)	**37.6**	**93.4**	**53.6**	56.4	<u>65.8</u>	<u>60.7</u>

Table 1. Method comparison for ZSL and GZSL tasks on MS-COCO. Precision (P), recall (R), and F1 score (F1) at Top-3 predictions per image are reported. The best results are bolded and the second-best results are underlined. * means reproduced results.

4. Experiments

4.1. Experimental Setup

Datasets. Following the split of seen and unseen classes in previous works [3, 16, 17], we evaluate on MS-COCO [22], NUS-WIDE [11], and Open Images (v4) [18]. In contrast, our method is trained only on texts without training images, and thus both seen and unseen classes can be considered as novel classes.

Metrics. As in the work [3], precision, recall, and F1 score at the Top-3 predictions per image are reported on the MS-COCO dataset. Similarly, following the works [3, 16, 17, 29], the mAP across all categories along with precision, recall, and F1 scores for the Top-3 and Top-5 predictions per image are presented on the NUS-WIDE dataset. Except for differences in the F1 scores for the Top-10 and Top-20 predictions per image, other evaluation metrics on Open Images are identical to those on NUS-WIDE.

Methods	Task	NUS-WIDE (#seen / #unseen = 925/81)							Open Images (#seen / #unseen = 7186/400)						
		Top-3			Top-5			mAP	Top-10			Top-20			mAP
		P	R	F1	P	R	F1		P	R	F1	P	R	F1	
CONSE [30]	ZSL	17.5	28.0	21.6	13.9	37.0	20.2	9.4	0.2	7.3	0.4	0.2	11.3	0.3	40.4
	GZSL	11.5	5.1	7.0	9.6	7.1	8.1	2.1	2.4	2.8	2.6	1.7	3.9	2.4	43.5
LabelEM [1]	ZSL	15.6	25.0	19.2	13.4	35.7	19.5	7.1	0.2	8.7	0.5	0.2	15.8	0.4	40.5
	GZSL	15.5	6.8	9.5	13.4	9.8	11.3	2.2	4.8	5.6	5.2	3.7	8.5	5.1	45.2
Fast0Tag [45]	ZSL	22.6	36.2	27.8	18.2	48.4	26.4	15.1	0.3	12.6	0.7	0.3	21.3	0.6	41.2
	GZSL	18.8	8.3	11.5	15.9	11.7	13.5	3.7	14.8	17.3	16.0	9.3	21.5	12.9	45.2
LESA (M=10) [17]	ZSL	25.7	41.1	31.6	19.7	52.5	28.7	19.4	0.7	25.6	1.4	0.5	37.4	1.0	41.7
	GZSL	23.6	10.4	14.4	19.8	14.6	16.8	5.6	16.2	18.9	17.4	10.2	23.9	14.3	45.4
BiAM [29]	ZSL	26.6	42.5	32.7	20.5	54.6	29.8	25.9	3.9	30.7	7.0	2.7	41.9	5.5	65.6
	GZSL	25.2	11.1	15.4	21.6	15.9	18.2	9.4	13.8	15.9	14.8	9.7	22.3	14.8	**81.7**
SDL (M=7) [3]	ZSL	24.2	41.3	30.5	18.8	53.4	27.8	25.9	6.1	47.0	10.7	4.4	68.1	8.3	62.9
	GZSL	27.7	13.9	18.5	23.0	19.3	21.0	12.1	35.3	40.8	37.8	23.6	54.5	32.9	75.3
$(ML)^2$-Enc [27]	ZSL	-	-	32.8	-	-	32.3	29.4	-	-	7.5	-	-	6.5	65.7
	GZSL	-	-	15.8	-	-	19.2	10.2	-	-	27.6	-	-	24.1	79.9
CLIP-FT [16]	ZSL	19.1	30.5	23.5	14.9	39.7	21.7	30.5	10.8	84.0	19.1	5.9	92.1	11.1	66.2
	GZSL	33.2	14.6	20.3	27.4	20.2	23.2	16.8	37.5	43.3	40.2	25.4	**58.7**	35.4	77.5
DualCoOp [37]	ZSL	37.3	46.2	41.3	28.7	59.3	38.7	43.6	-	-	-	-	-	-	-
	GZSL	31.9	13.9	19.4	26.2	19.1	22.1	12.0	-	-	-	-	-	-	-
MKT [16]	ZSL	27.7	44.3	34.1	21.4	57.0	31.1	37.6	11.1	86.8	19.7	6.1	94.7	11.4	68.1
	GZSL	**35.9**	**15.8**	**22.0**	**29.9**	**22.0**	**25.4**	**18.3**	**37.8**	**43.6**	**40.5**	**25.4**	58.5	**35.4**	81.4
TagCLIP* [23]	ZSL	31.4	39.0	34.8	26.0	53.9	35.1	40.1	7.2	56.1	12.8	4.7	72.8	8.8	32.0
	GZSL	24.8	10.8	15.1	20.3	14.7	17.0	12.7	12.6	14.5	13.5	9.1	20.9	12.7	24.2
CoMC [26]	ZSL	33.5	53.5	41.2	24.8	66.1	36.1	48.2	-	-	-	-	-	-	-
	GZSL	-	-	-	-	-	-	-	-	-	-	-	-	-	-
RCNn(ours)	ZSL	**43.7**	**54.3**	**48.4**	**33.6**	**69.6**	**45.3**	**53.3**	**12.0**	**93.3**	**21.2**	**6.3**	**98.0**	**11.8**	**70.2**
	GZSL	31.8	13.8	19.3	26.7	19.4	22.4	17.9	23.3	26.8	24.9	16.8	38.6	23.4	79.2

Table 2. Comparison between our method and the state-of-the-art methods for ZSL and GZSL tasks on NUS-WIDE and Open Images. The results of mAP over all classes, as well as precision (P), recall (R), and F1 score (F1) are reported. Top-3 and Top-5 predictions for NUS-WIDE and Top-10 and Top-20 predictions for Open Images in each image are used to compute P, R, and F1. The best results are marked in bold, and the second-place results are underlined. * means reproduced results.

All methods are evaluated under both the ZSL task and the GZSL task. More experimental settings are provided in Section C of the supplementary material.

4.2. Comparison with State-of-the-Arts

The vision-driven methods include CONSE [30], LabelEM [1], Fast0Tag [45], Deep0Tag [34], LESA [17], BiAM [29], SDL [3], $(ML)^2$-Enc [27], CLIP-FT [16], DualCoOp [37] and MKT [16]. All methods belong to the separate alignment other than CLIP-FT, DualCoOp, and MKT. CLIP-FT, DualCoOp, and MKT fine-tune the image encoder or contextual prompts for CLIP via image data. TagCLIP [23] directly utilizes the transfer capabilities of the original CLIP without any extra training. The language-driven CoMC [26] only uses text data for training. The above methods serve as the baseline for comparison with our method, and the results are shown on three datasets.

Performance on MS-COCO. Tab. 1 shows the performance comparison between our method and the state-of-the-art methods on MS-COCO. Our method achieves the best results for each metric in the ZSL task, which is 2.3%, 5.8%, and 3.3% higher than the second-best DualCoOp in precision, recall, and F1 score, respectively. Most of the categories in GZSL are seen classes, which are trained with image data in the vision-driven methods. Ideally, the performance of those methods has an advantage in the GZSL task. In contrast, our method treats the seen classes as novel classes, yet still achieves competitive performance. Our recall and F1 scores are surpassed only by those of the vision-driven DualCoOp, outperforming all other vision-driven methods. Although the training-free TagCLIP also treats the seen classes as novel classes like ours, our approach significantly outperforms TagCLIP by more than 10% in three metrics for the GZSL task.

Performance on NUS-WIDE. As shown in Tab. 2, our method achieves the best performance of all metrics in ZSL. Compared with the second-best results, F1 score improves by 7.1% @ Top-3 and 6.6% @ Top-5, and mAP improves by 5.1%. Our method has obvious advantages over both language-driven CoMC and vision-driven DualCoOp. In the GZSL task, our method can achieve comparable performance on NUS-WIDE without training on image data. Our mAP is only 0.4% lower than the highest MKT. Compared with the vision-driven DualCoOp, our F1 score @ Top-5 and mAP improve by 0.3% and 5.9%, respectively. In addition, our F1 @ Top-3, F1 @ Top-5, and mAP exceed the

Baseline	CNR	PBL	MSSA	Task	MS-COCO mAP	MS-COCO F1 (Top-3)	MS-COCO F1 (Top-5)	NUS-WIDE mAP	NUS-WIDE F1 (Top-3)	NUS-WIDE F1 (Top-5)
✓				ZSL	74.6	48.0	35.1	46.5	38.4	35.5
				GZSL	57.0	46.4	41.2	16.3	18.2	20.8
	✓			ZSL	82.2	52.2	37.5	51.7	45.9	43.2
				GZSL	66.7	54.8	49.0	17.7	19.5	22.5
	✓	✓		ZSL	84.1	52.8	37.4	51.7	47.4	44.4
				GZSL	69.6	59.4	51.0	17.9	19.5	22.5
	✓	✓	✓	ZSL	86.3	53.6	37.7	53.3	48.4	45.3
				GZSL	73.1	60.7	52.0	17.9	19.3	22.4

Table 3. Ablation study on the main components of our method. For the ZSL and GZSL tasks, mAP over all classes and F1 scores of Top-3 and Top-5 predictions on MS-COCO and NUS-WIDE are reported.

Figure 3. Prediction comparison on test samples from NUS-WIDE. GT means ground truth labels. Top-5 (top row) and Top-10 (bottom row) predictions are shown for ZSL and GZSL, respectively. The green, red, and black fonts denote true positive predictions, incorrect predictions, and reasonable predictions.

performance of all separate alignment methods (above the middle dividing line in Tab. 2).

Performance on Open Images. For the ZSL task, in Tab. 2, our F1 @ Top-10, F1 @ Top-20, and mAP are 1.5%, 0.4% and 2.1% higher than those of MKT respectively. Compared with SDL (M=7), our F1 @ Top-10, F1 @ Top-20, and mAP improve by 10.5%, 3.5%, and 7.3%. In the GZSL task, our method achieves a higher F1 score than CONSE, LabelEM, Fast0tag, LESA, BiAM, and TagCLIP. The vision-driven MKT and CLIP-FT benefit from seen classes with at least 100 images each, providing sufficient training data and leading to significant F1 scores. Despite the absence of training images, our method still delivers competitive performance, achieving an mAP that surpasses CLIP-FT by 1.7%.

4.3. Ablation Study

We encode prompts with the original class names as class embeddings to match visual features in CLIP, establishing our baseline. The key components compared to this baseline include class name reconstruction (CNR), pair-based loss (PBL), and multi-snippet semantic aggregation (MSSA). As shown in Tab. 3, compared with the baseline, the prompts containing reconstructed class names are encoded into class embeddings for the ZSL and GZSL tasks, which significantly improves mAP, F1 @ Top-3, and F1 @ Top-5 on both two datasets. For MS-COCO, both seen and unseen classes are objects, with no inherent scene or semantic hierarchy. Extracting visual features, semantic hierarchies, and co-occurrence scenes can enhance class name semantics and improve classification performance. PBL strengthens the distinction between reconstructed class names to facilitate better recognition, especially F1 based on ranking predictions of each image in GZSL. MSSA further improves performance across all metrics.

For NUS-WIDE, we can find that PBL and MSSA can not improve the performance of GZSL. Seen and unseen classes contain rich semantic granularity (see section A of supplementary material for details), and seen classes without human validation have too many synonyms (Fig. 3 bottom, such as clouds, cloud, cloudy). The textual descriptions of a class include the hierarchical relationship with its superclasses, and the class and its synonyms often share highly similar descriptions. Consequently, CNR strongly links it to its superclasses and synonyms. This association is difficult to penalize using PBL. As shown in Fig. 3, a class, along with its synonyms and superclasses, is often predicted simultaneously in each image for GZSL, which is reasonable. However, the ground truth for GZSL consists of a diverse set of classes with only a few superclasses and synonyms. As a result, ranking-based evaluation metrics such as mAP and F1 struggle to accurately assess the predictions. MSSA utilizes more detailed features that are more friendly to predicting its synonyms and superclasses. This may further reduce the diversity of the top-ranked prediction classes, resulting in a decrease in performance. For the

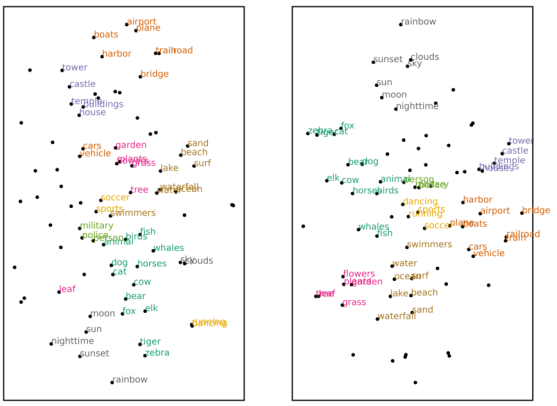

Figure 4. t-SNE visualization of class embeddings based on original class names (left) and reconstructed class names (right).

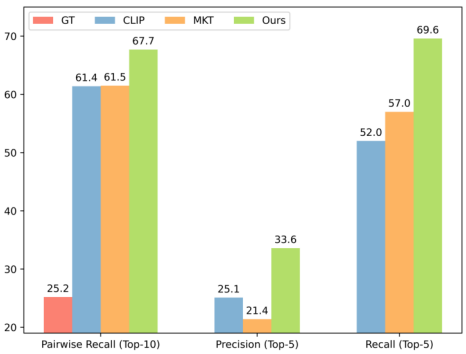

Figure 5. Pairwise recall at Top-10 predictions, along with label-level precision and recall at Top-5 predictions, for CLIP, MKT, and our method on the NUS-WIDE test set.

ZSL task, unseen classes include objects, scenes, and superclasses but lack synonyms. Since most classes have fundamentally distinct semantics, PBL is prone to impose dissimilarity penalties. MSSA provides more discriminative semantic information for recognition. Thus, PBL makes F1 @ Top-3 and Top-5 improve by 1.5% and 1.2% , and MSSA also boosts them by 1.0% and 0.9%.

4.4. Qualitative Evaluation

In Fig. 3, compared to CLIP and MKT, our predictions align more closely with the GT in ZSL, which is also reflected in the higher F1 scores of our method. Additionally, our predictions better capture semantic granularity relationships, such as "sunset" and "sun" or "clouds" and "sky." For GZSL, the GT and predictions of MKT and CLIP contain a wider variety of classes than ours. However, our method prioritizes predicting synonyms and superclasses of the top-ranked categories in the outputs, ensuring semantic consistency. This highlights the effectiveness of our approach in handling semantic granularity issues.

4.5. Semantic Granularity Analysis

Qualitative Analysis. To illustrate the semantic relationships between reconstructed class names, prompts with reconstructed class names are encoded into class embeddings for t-SNE visualization. The class embeddings of the original class names are also visualized for comparison. We select some categories from the 81 unseen classes of NUS-WIDE to create eight semantically related groups, namely animal, person, vehicle, buildings, plants, sports, water, and sky. These group names are also part of the unseen classes. As shown in Fig. 4, compared with class embeddings of original class names, the class embedding distribution of the reconstructed class names is tighter in each group. For example, "leaf" and "plants", "sun" and "sky", as well as "dancing" and "sports" are closer. Meanwhile, "fish" is located between "water" and "animal", and "swimmers" is located between "water" and "sports".

Quantitative Analysis. To further validate the effectiveness of our approach in handling the semantic granularity problem, we present quantitative results in Fig. 5. Pairwise recall is proposed to measure the recognition ability of true hierarchical pairs (*e.g.*, cat-animal) in the predictions. It is the proportion of true hierarchical pairs in the predictions to all hierarchical pairs in the ground truth (GT). We find that the parent classes of many categories are often overlooked in the ground truth. Therefore, we complement the parent classes and count the hierarchical pairs for the ground truth, based on the hierarchical relationships within the eight groups above. We can see that the number of hierarchical pairs in GT before complement is very limited in Fig. 5. Compared with CLIP and MKT, our method achieves higher pairwise recall when maintaining high label-level precision and recall. This means that our method is better at predicting hierarchical pairs.

5. Conclusion

In this work, we propose a novel language-driven zero-shot framework for multi-label classification with semantic granularity. To this end, we prompt a large language model to generate class-related text descriptions, including visual attributes, hierarchical relationships, and co-occurrence scenes for each class. Then, the collected text descriptions are utilized to reconstruct class names. Additionally, we impose a pair-based loss to enhance the distinctiveness of reconstructed class names. During inference, we aggregate semantic predictions from multiple image snippets at both global and local levels. Extensive experimental results demonstrate that our language-driven method achieves state-of-the-art performance in the zero-shot task and surpasses most image-trained methods in the generalized zero-shot task. The results highlight the effectiveness of our approach in cross-granularity classification.

Acknowledgements. The work was supported by the National Key R&D Program of China under Grant 2021ZD0201300, the National Natural Science Foundation of China under Grant 623B2040, the Foundation for Outstanding Research Groups of Hubei Province of China under Grant 2025AFA012, and the 111 Project on Computational Intelligence and Intelligent Control under Grant B18024.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(7):1425–1438, 2015. 6

[2] Björn Barz and Joachim Denzler. Hierarchy-based image embeddings for semantic image retrieval. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 638–647. IEEE, 2019. 3

[3] Avi Ben-Cohen, Nadav Zamir, Emanuel Ben-Baruch, Itamar Friedman, and Lihi Zelnik-Manor. Semantic diversity learning for zero-shot multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 640–650, 2021. 1, 2, 5, 6

[4] Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12506–12515, 2020. 3

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. 2, 3, 4

[6] Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. An empirical study on large-scale multi-label text classification including few and zero-shot labels. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7503–7515, 2020. 3

[7] Shiming Chen, Guosen Xie, Yang Liu, Qinmu Peng, Baigui Sun, Hao Li, Xinge You, and Ling Shao. Hsva: Hierarchical semantic-visual adaptation for zero-shot learning. *Advances in Neural Information Processing Systems*, 34:16622–16634, 2021. 3

[8] 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. 1, 2

[9] 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*, pages 339–346, 2022. 2

[10] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2019. 1, 2

[11] 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. 1, 2, 5

[12] Jia Deng, Alexander C Berg, Kai Li, and Li Fei-Fei. What does classifying more than 10,000 image categories tell us? In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part V 11*, pages 71–84. Springer, 2010. 3

[13] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in Neural Information Processing Systems*, 26, 2013. 3

[14] Yunhao Ge, Jie Ren, Andrew Gallagher, Yuxiao Wang, Ming-Hsuan Yang, Hartwig Adam, Laurent Itti, Balaji Lakshminarayanan, and Jiaping Zhao. Improving zero-shot generalization and robustness of multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11093–11101, 2023. 3

[15] Zixian Guo, Bowen Dong, Zhilong Ji, Jinfeng Bai, Yiwen Guo, and Wangmeng Zuo. Texts as images in prompt tuning for multi-label image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2808–2817, 2023. 1, 2

[16] Sunan He, Taian Guo, Tao Dai, Ruizhi Qiao, Xiujun Shu, Bo Ren, and Shu-Tao Xia. Open-vocabulary multi-label classification via multi-modal knowledge transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 808–816, 2023. 1, 2, 3, 5, 6

[17] Dat Huynh and Ehsan Elhamifar. A shared multi-attention framework for multi-label zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8776–8786, 2020. 1, 2, 5, 6

[18] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 1, 2, 5

[19] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *European Conference on Computer Vision*, pages 143–160. Springer, 2024. 4

[20] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 16478–16488, 2021. 1, 2

[21] Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Multi-label zero-shot learning with

structured knowledge graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1576–1585, 2018. 1, 2

[22] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5

[23] Yuqi Lin, Minghao Chen, Kaipeng Zhang, Hengjia Li, Mingming Li, Zheng Yang, Dongqin Lv, Binbin Lin, Haifeng Liu, and Deng Cai. Tagclip: A local-to-global framework to enhance open-vocabulary multi-label classification of clip without training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3513–3521, 2024. 5, 6

[24] Hui Liu, Danqing Zhang, Bing Yin, and Xiaodan Zhu. Improving pretrained models for zero-shot multi-label text classification through reinforced label hierarchy reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1051–1062, 2021. 3

[25] Mingxuan Liu, Tyler L Hayes, Elisa Ricci, Gabriela Csurka, and Riccardo Volpi. Shine: Semantic hierarchy nexus for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16634–16644, 2024. 4

[26] Yicheng Liu, Jie Wen, Chengliang Liu, Xiaozhao Fang, Zuoyong Li, Yong Xu, and Zheng Zhang. Language-driven cross-modal classifier for zero-shot multi-label image recognition. In *Forty-first International Conference on Machine Learning*, 2024. 3, 6

[27] Ziming Liu, Song Guo, Xiaocheng Lu, Jingcai Guo, Jiewei Zhang, Yue Zeng, and Fushuo Huo. Ml^{2}p-encoder: On exploration of channel-class correlation for multi-label zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23859–23868, 2023. 6

[28] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *the Eleventh International Conference on Learning Representations*, 2023. 4

[29] Sanath Narayan, Akshita Gupta, Salman Khan, Fahad Shahbaz Khan, Ling Shao, and Mubarak Shah. Discriminative region-based multi-label zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8731–8740, 2021. 1, 2, 3, 5, 6

[30] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013. 5, 6

[31] Zachary Novack, Julian McAuley, Zachary Chase Lipton, and Saurabh Garg. Chils: Zero-shot image classification with hierarchical label sets. In *International Conference on Machine Learning*, pages 26342–26362. PMLR, 2023. 3

[32] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. 1

[33] 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, 3

[34] Shafin Rahman, Salman Khan, and Nick Barnes. Deep0tag: Deep multiple instance learning for zero-shot image tagging. *IEEE Transactions on Multimedia*, 22(1):242–255, 2019. 5, 6

[35] Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben-Baruch, and Asaf Noy. Ml-decoder: Scalable and versatile classification head. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 32–41, 2023. 2

[36] Miguel E Ruiz and Padmini Srinivasan. Hierarchical text categorization using neural networks. *Information Retrieval*, 5:87–118, 2002. 3

[37] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. In *Advances in Neural Information Processing Systems*, pages 30569–30582, 2022. 1, 2, 5, 6

[38] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2016. 1, 2

[39] Shouwen Wang, Qian Wan, Xiang Xiang, and Zhigang Zeng. Saliency regularization for self-training with partial annotations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1611–1620, 2023. 2

[40] Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. Multi-label classification with label graph superimposing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12265–12272, 2020. 1, 2

[41] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 464–472, 2017. 1, 2

[42] Feihong Wu, Jun Zhang, and Vasant Honavar. Learning classifiers using hierarchically structured class taxonomies. In *International Symposium on Abstraction, Reformulation, and Approximation*, pages 313–320. Springer, 2005. 3

[43] Vacit Oguz Yazici, Abel Gonzalez-Garcia, Arnau Ramisa, Bartlomiej Twardowski, and Joost van de Weijer. Orderless recurrent models for multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13440–13449, 2020. 1, 2

[44] Kai Yi, Xiaoqian Shen, Yunhao Gou, and Mohamed Elhoseiny. Exploring hierarchical graph representation for large-scale zero-shot image classification. In *European Conference on Computer Vision*, pages 116–132. Springer, 2022. 3

[45] Yang Zhang, Boqing Gong, and Mubarak Shah. Fast zero-shot image tagging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5985–5994, 2016. 1, 2, 5, 6

[46] Jiawei Zhao, Ke Yan, Yifan Zhao, Xiaowei Guo, Feiyue Huang, and Jia Li. Transformer-based dual relation graph for multi-label image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 163–172, 2021. 1, 2