Unleashing the Potential of Hierarchical Region Clues for Open-Vocabulary Multi-Label Classification

Peirong Ma, Wu Ran, Zhiquan He, Jian Pu, *Member, IEEE*, and Hong Lu, *Member, IEEE*

***Abstract*—Open-vocabulary multi-label classification (OV-MLC) aims to leverage the rich multi-modal knowledge from Vision-language pre-training (VLP) models to further improve the recognition ability for unseen (novel) classes beyond the training set in multi-label scenarios. Existing OV-MLC methods only perform predictions on single hierarchical regions, and aggregate the prediction scores of these regions through simple *top-k* mean pooling. This fails to unleash the potential of rich hierarchical region clues in multi-label images and does not fully exploit the discriminative information from all regions in the image, resulting in sub-optimal performance. In this work, we propose a novel OV-MLC framework to fully harness the power of multiple hierarchical region clues. Specifically, we first design a hierarchical clue gathering (HCG) module to gather different hierarchical clues, enabling more precise recognition of multiple object categories with different sizes in a multi-label image. Then, by viewing multi-label classification as single-label classification of each region within the image, we present a novel hierarchical score aggregation (HSA) approach, thereby better utilizing the predictions of each image region for each class. We also utilize a well-designed region selection strategy (RSS) to eliminate noise or background regions in an image that are irrelevant to classification, achieving higher multi-label classification accuracy. In addition, we propose a hybrid prompt learning (HPL) strategy to enhance visual-semantic consistency while preserving the generalization capability of label embeddings for unseen classes. Extensive experiments on public benchmark datasets demonstrate that our method significantly outperforms the current state-of-the-art.**

***Index Terms*—Open-vocabulary multi-label classification, zero-shot multi-label classification, vision-language pre-training model, hierarchical region clues.**

I. Introduction

MULTI-LABEL classification (MLC) [1], [2], [3], [4] is one of the most extensively studied tasks in computer vision, aiming to recognize multiple objects, scenes, or concepts present in an image. With the adoption of deep learning, MLC methods [5], [6], [7], [8], [9] have made considerable progress. However, in the traditional MLC task, the candidate label set in the training phase and the testing phase is the same, which is far from meeting the requirements of some real-world applications because it does not have the ability to identify emerging unseen (novel) classes. To address this issue, Zero-shot multi-label classification (ZS-MLC) [10], [11], [12], [13], as a cross-task of zero-shot learning (ZSL) [14], [15], [16], [17], [18] and MLC, has attracted increasing attention in recent years. ZS-MLC aims to predict multiple unseen class labels in a multi-label image, primarily achieved by transferring knowledge from seen classes to unseen classes via a language model (e.g. GloVe [19]) pre-trained on large-scale corpora. However, this approach only explores knowledge transfer in text modality and ignores knowledge transfer in visual modality and cross-modality, thereby severely limiting further performance improvement.

Open-vocabulary multi-label classification (OV-MLC) [20], [21] aims to utilize the multi-modal knowledge from Vision-language pre-training (VLP) models (e.g., CLIP [22]) to further improve the recognition performance of unseen classes in multi-label scenarios, it can be regarded as a relaxed/special version of ZS-MLC. Specifically, OV-MLC assumes that in addition to the annotated data of the seen classes, a low-cost auxiliary supervised data source is accessible during training (e.g., the large-scale image-text dataset for pre-training VLP models is crawled from the web, requiring no laborious manual annotation and easily available). The VLP model connects visual concepts to text descriptions and learns an unbounded (open) visual concept vocabulary from large-scale image-text pairs. We can generate label semantic embeddings of arbitrary concepts (categories) through the text encoder of the VLP model, which makes the label prediction for open-vocabulary possible. Due to the ability to expand recognition vocabulary based on rich multi-modal knowledge within VLP models, OV-MLC is more general, practical, and effective compared to ZS-MLC.

Although VLP models exhibit remarkable zero-shot transfer capabilities on single-label image classification, and extracting knowledge from off-the-shelf VLP models has also become a popular solution for OV-based tasks (such as open-vocabulary object detection [23], [24], [25], [26], [27] or open-vocabulary

Received 12 March 2024; revised 8 January 2025; accepted 29 March 2025. Date of publication 6 October 2025; date of current version 17 December 2025. This work was supported by the National Natural Science Foundation of China under Grant 62072112. The associate editor coordinating the review of this article and approving it for publication was Dr. Yong Luo. *(Corresponding authors: Jian Pu; Hong Lu.)*

Peirong Ma, Wu Ran, Zhiquan He, and Hong Lu are with the Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai 200438, China (e-mail: prma20@fudan.edu.cn; wran21@m.fudan.edu.cn; zqhe22@m.fudan.edu.cn; honglu@fudan.edu.cn).

Jian Pu is with the Institute of Science and Technology for Brain-inspired Intelligence, Fudan University, Shanghai 200438, China (e-mail: jianpu@fudan.edu.cn).

Digital Object Identifier 10.1109/TMM.2025.3618542

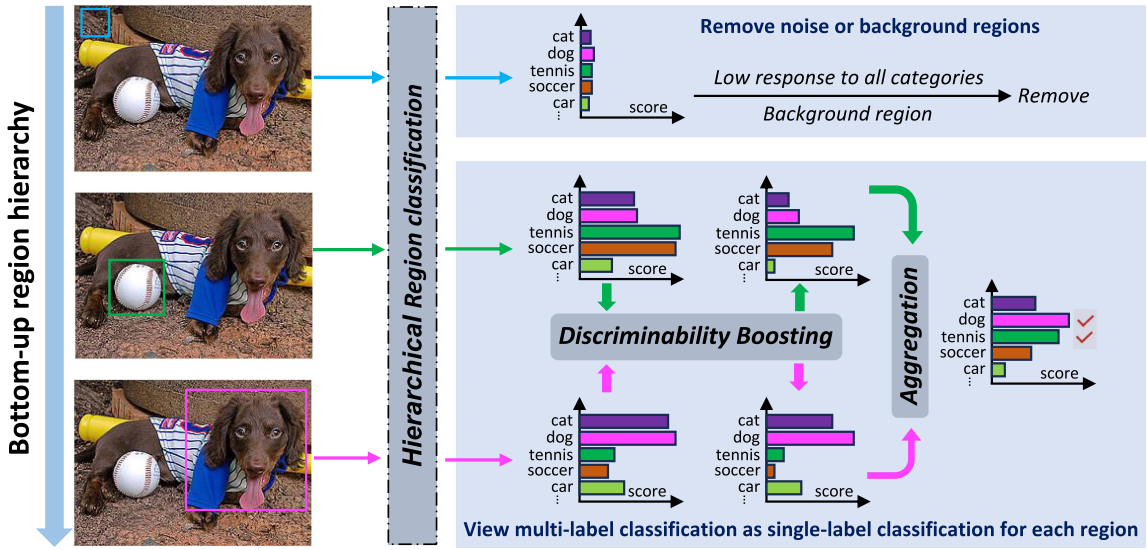


Fig. 1. Overall idea of the proposed HRCP. We gather region clues across various hierarchies to fully unleash their potential for recognizing multiple object categories of varying sizes in an image. For example, the smaller object class '*tennis*' and the larger object class '*dog*' are recognized at different region hierarchies. Moreover, for the prediction scores of each region, we first perform discriminability boosting and then aggregate these enhanced region scores as the final region-based predictions. Additionally, we further improve the classification accuracy by removing noise or background regions that are low-responsive to all categories.

semantic segmentation [28], [29], [30], [31], [32]), but transferring the zero-shot classification capabilities of VLP models to OV-MLC is still a challenging research direction. Because these VLP models are typically trained by aligning global image embeddings with text embeddings, their outputs only focus on the scene-level global information of the image. Therefore, they only capture dominant labels, ignoring labels associated with smaller regions and lacking the ability to recognize multiple categories within an image. Compared with global information, local (region) information is more crucial for multi-label classification tasks. Therefore, existing works on OV-MLC [20], [21] all exploit the local information of the image and achieve significant success by exploring the similarity between image regions and text labels. MKT [20] is currently the state-of-the-art method for OV-MLC. However, it still suffers from three main issues. Firstly, MKT only captures single-level region clues, limiting its ability to effectively recognize object categories of various sizes in multi-label images. Second, same as the classic ZS-MLC method BiAM [12], MKT first classifies single-level region features, and then performs the spatial *top-k* mean pooling on their class predictions to obtain a region-based prediction score. However, this simple averaging of the *top-k* predictions for each class across all regions is sub-optimal and fails to fully exploit the discriminative information of all regions in an image. Third, the prompt tuning in MKT obtains better prompts by optimizing the token embedding layer of CLIP's text encoder on seen labels, which leads to biased model. Because fine-tuning the token embedding layer destroys the strong representation and generalization ability of the CLIP's text encoder, and it is also easy to overfit on the seen classes, especially when the dataset contains only a small number of seen labels. Overall, a more effective paradigm for OV-MLC needs to be established to fully explore the multi-modal knowledge from VLP models and the rich hierarchical clues in multi-label images.

Based on the above analysis, we propose a novel framework named HRCP to fully unleash the **P**otential of **H**ierarchical **R**egion **C**ues in multi-label images. As depicted in Fig. 1, we first gather different hierarchical region clues to facilitate the recognition of multiple object categories across various scales within an image. Then, we treat multi-label classification as the single-label classification for each region in the image, and propose a novel way to aggregate hierarchical region scores to replace spatial *top-k* mean pooling. At test time, we also design a novel region selection strategy, which further improves performance by filtering out noise or background regions of the image that are not relevant for classification. Additionally, to preserve the generalization ability of label embeddings extracted by the VLP text encoder while enhancing visual-semantic consistency, we propose to combine learnable prompt and fixed prompt, as illustrated in Fig. 2. Compared with fine-tuning the token embedding layer in MKT, this hybrid prompt learning approach significantly improves the classification accuracy, and requires less computing resources, with faster training speed.

The contributions can be summarized as follows:

1) We propose HRCP, a novel Open-vocabulary multi-label classification (OV-MLC) framework, which fully harnesses the power of hierarchical region clues to more accurately recognize multiple object categories with different sizes in an image.
2) We propose a novel hierarchical score aggregation (HSA) approach to better utilize the predictions of each image region for each class. At test time, we also design a region selection strategy (RSS), which can further improve the prediction accuracy by removing classification-independent noise or background regions in images.
3) We propose a hybrid prompt learning strategy to make the pre-trained CLIP text encoder can better adapt to the OV-MLC task. This hybrid prompt learning method not

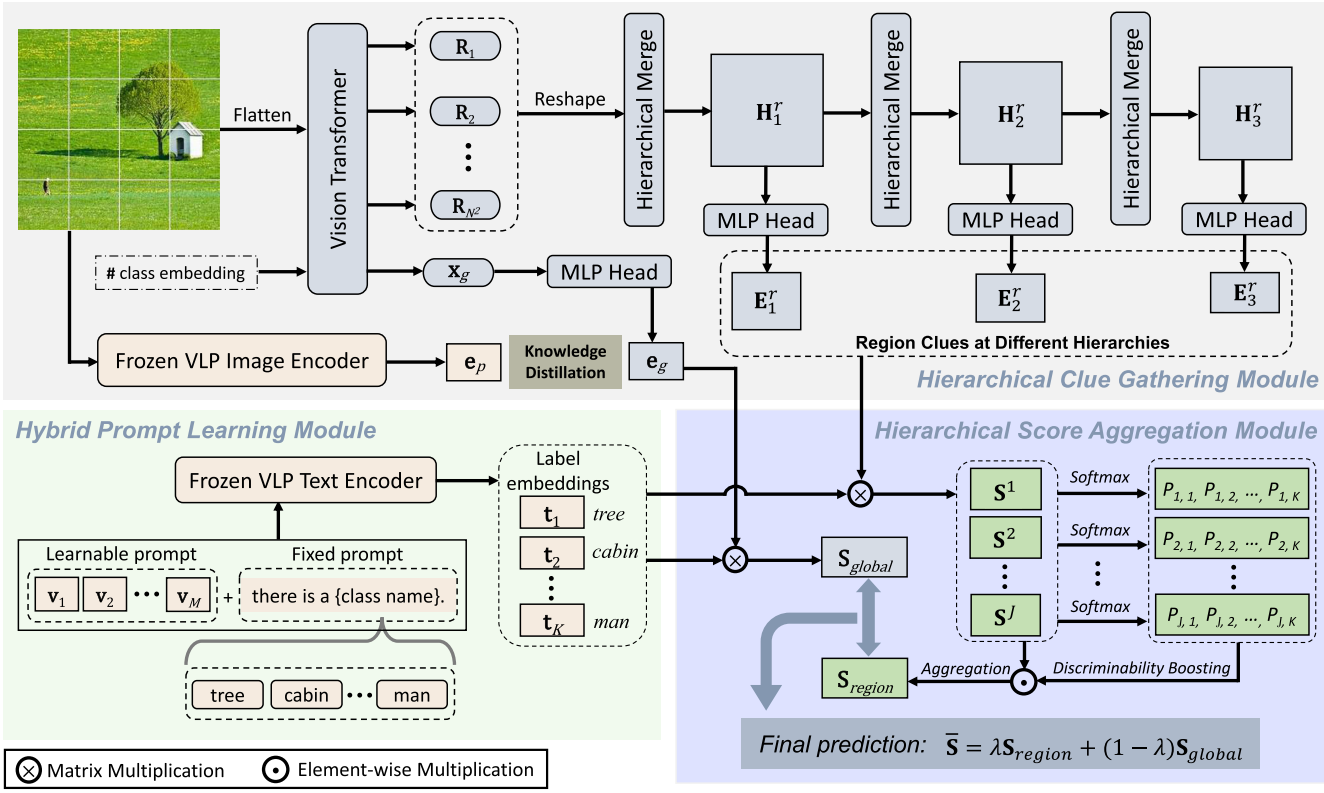


Fig. 2. The architecture of the proposed HRCP, which consists of a hierarchical clue gathering (HCG) module, a hierarchical score aggregation (HSA) module, and a hybrid prompt learning (HPL) module.

only improves the visual-semantic consistency, but also maintains the generalization ability of the generated label embeddings for unseen classes.

4) Experiments on two large-scale OV-MLC benchmark datasets demonstrate that the proposed HRCP achieves new state-of-the-art performance.

The rest of this paper is organized as follows. We first review the related work in Section II. Then, Section III provides a detailed description of the proposed method. Subsequently, Section IV presents the experimental setup and reports the experimental results. Finally, Section V concludes this paper.

II. Related Work

A. Zero-Shot Multi-Label Classification

Zero-shot multi-label classification (ZS-MLC) methods [10], [11], [12], [13] train on annotated seen classes and improve their generalization ability to unseen classes beyond the training set by exploiting the semantic correlation between seen and unseen labels. For instance, Fast0tag [33] and SDL [11] train a network to estimate the principal direction of an image, so that the word vectors of relevant labels are ranked ahead of irrelevant labels. SKG [34] learns a label propagation mechanism from the semantic space using a knowledge graph, enabling the reasoning of the model for predicting unseen labels. LESA [10] introduces a shared multi-attention framework and BiAM [12] proposes a bi-level attention module to recognize multiple seen and unseen labels in an image based on attention mechanisms. Gen-MLZSL [35] proposes three different fusion methods to generate multi-label features for unseen classes. $(ML)^2P$-Encoder [13] extracts and preserves channel-wise semantics by exploring the channel-class correlation. However, In these ZS-MLC methods, the backbone networks for extracting image features (e.g., VGG [36]) and label embeddings (e.g., GloVe [19]) are independently designed and trained separately. As a result, the extracted image features and label embeddings are in different data spaces, and the visual-semantic consistency between them is low, thereby resulting in poor performance of these methods.

B. Vision-Language Pre-Training Models

Recently, many Vision-language pre-training (VLP) models such as CLIP [22], ALIGN [37] and SLIP [38] are proposed, which aim to match image embeddings with text embeddings in a cross-modal common semantic space. These models are trained on large-scale (billions of) image-text pairs collected from the web and obtain powerful image-text representation capabilities through contrastive learning [39], [40]. After pre-training, VLP models can zero-shot transfer to downstream tasks, endowing them with linguistic capabilities. For example, in classification tasks, when a novel class emerges, we can synthesize the class label embedding by inputting the natural language description of that class (e.g.,"A photo of a {class name}.") into the VLP

text encoder. Then, we compute the similarity between the image embedding generated by the image encoder and this label embedding as the prediction result.

Benefiting from cross-modal pre-training on large-scale image-text pairs, VLP models have acquired multi-modal knowledge of general concepts and achieved impressive results on the ZS-MLC task. This suggests that text embeddings learned jointly with visual data can better encode visual similarities between concepts than label embeddings learned from language corpora alone (e.g., GloVe [19]). However, these VLP models are trained based on global image-text feature alignment, and the learned knowledge is limited to scene-level global image representation. Global representations capture the overall contextual information of an image, often dominated by the most prominent/common/larger categories or concepts in the image, which is highly effective for single-label classification. However, it ignores some less conspicuous objects, such as smaller-sized object categories, hence not suitable for multi-label classification tasks that require identifying multiple object categories of different sizes within an image. In multi-label settings, discriminative local region features are more helpful.

C. Open-Vocabulary Multi-Label Classification

The Open-vocabulary setting is first introduced by Zareian et al. [41] for object detection tasks. This is a special zero-shot learning setting. It assumes that in addition to annotated seen class data, a VLP model pre-trained on large-scale image-text pairs can be utilized. In this setting, while unseen classes are unknown during training, they can belong to any subset of the entire language vocabulary in the pre-training task (e.g., contrastive learning on large-scale image-text pairs). This setup has proven to be very effective in some computer vision tasks such as object detection [23], [24], [25], [26], [27] and semantic segmentation [28], [29], [30], [31], [32]). Open-vocabulary multi-label classification (OV-MLC) is a novel emerging subfield of zero-shot learning derived in the context of large-scale VLP models. It aims to transfer rich multi-modal knowledge from VLP models, thereby further improving the recognition performance of unseen labels in multi-label scenarios. There are a limited number of studies on OV-MLC. OVML-VLP [21] introduces an image-text attention module after CLIP's image encoder and proposes a contrastive loss training method to help the attention module better utilize image features from different regions. MKT [20] transfers knowledge from the image encoder of the VLP model through knowledge distillation, and enables label embeddings generated by the text encoder of the VLP model to better support the OV-MLC task through prompt tuning.

III. The Proposed Method

A. Problem Definition

Following previous work on OV-MLC [20] and ZS-MLC [10], assume the entire label set: $\mathcal{Y} = \mathcal{Y}^S \cup \mathcal{Y}^U$, where $\mathcal{Y}^S$ denotes the seen label set for training, $\mathcal{Y}^U$ denotes the unseen label set without training images, and $\mathcal{Y}^S \cap \mathcal{Y}^U = \phi$. Let $\{(x_m, y_m); m = 1, 2, \ldots, M\}$ denote M multi-label training images, where x_m denotes the m-th image and $y_m \subseteq \mathcal{Y}^S$ denotes the set of seen labels present in this image. The corresponding label embedding $\{\mathbf{t}_y\}_{y\in\mathcal{Y}}$ for each class is obtained by feeding the class name into a pre-trained text encoder. Zero-shot learning (ZSL) aims to assign relevant unseen labels $y_i \subseteq \mathcal{Y}^U$ for a given test image x_i; while the more realistic and challenging Generalized zero-shot learning (GZSL) aims to assign relevant seen and unseen labels $y_i \subseteq \mathcal{Y}$ to a given test image x_i.

B. Overview

Fig. 2 illustrates the overall framework of our HRCP, which contains a hierarchical clue gathering (HCG) module, a hierarchical score aggregation (HSA) module, and a hybrid prompt learning (HPL) module. Specifically, HCG aims to gather different hierarchical region clues and scene-level global clues to learn to recognize multiple objects of different sizes in an image. By treating multi-label classification as the single-label classification for each region in an image, HSA first performs discriminative enhancement on hierarchical region prediction scores and then aggregates them to obtain region-based predictions. HPL combines learnable prompt and fixed prompt to improve visual-semantic consistency while preserving the generalization of the generated label embeddings to unseen classes. For a fair comparison with previous method [20], we use pre-trained CLIP [22] as our VLP model, which contains an image encoder to transfer scene-level global information and a text encoder to generate label embedding for each category.

C. Hierarchical Clue Gathering

Scene-level Global Clues: As shown in Fig. 2, the first component of the hierarchical clue gathering module is a standard Vision Transformer (ViT) [42], whose purpose is to gather scene-level global clues and single hierarchical region clues of an image. Specifically, the input image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ is first reshaped into a sequence of flattened 2D patches $\mathbf{x}_p \in \mathbb{R}^{N^2\times(P^2\cdot C)}$, where $H \times W$ is the resolution of the image and C is the channel number, $P \times P$ is the size of each patch, and $N^2 = HW/P^2$ is the total number of patches. Then, $\mathbf{x}_p$ is mapped to D-dimensional patch embeddings by a trainable linear projection $\mathbf{E}$. Subsequently, the processing of the ℓ-th block in the transformer encoder can be formulated as:

$$\begin{aligned}
\mathbf{h}_0 &= \left[\mathbf{x}_{\text{cls}};\ \mathbf{x}_p^1\mathbf{E};\ \mathbf{x}_p^2\mathbf{E};\ \cdots;\ \mathbf{x}_p^{N^2}\mathbf{E}\right] + \mathbf{E}_{\text{pos}}\,,\\
\mathbf{h}'_\ell &= \mathbf{h}_{\ell-1} + \text{MSA}\left(\text{LN}\left(\mathbf{h}_{\ell-1}\right)\right),\\
\mathbf{h}_\ell &= \mathbf{h}'_\ell + \text{MLP}\left(\text{LN}\left(\mathbf{h}'_\ell\right)\right),
\end{aligned} \tag{1}$$

where $\mathbf{x_{cls}}$ is the learnable class embedding, and $\mathbf{E_{pos}}$ is the position embedding. $\text{MSA}(\cdot)$, $\text{LN}(\cdot)$ and $\text{MLP}(\cdot)$ denote multi-head self-attention, layernorm and multilayer perceptron, respectively. In this work, the output of ViT is written as:

$$\mathbf{O}_L = [\mathbf{x}_g;\ \mathbf{R_1};\ \mathbf{R_2};\ \cdots;\ \mathbf{R}_{N^2}]\,, \tag{2}$$

where $\mathbf{x}_g$ is the output corresponding to $\mathbf{x_{cls}}$, which can be regarded as the scene-level global representation of the image.

$[\mathbf{R_1};\ \mathbf{R_2};\ \cdots;\ \mathbf{R}_{N^2}]$ are the output corresponding to each image patch, and each of them can be regarded as a region representation of size $P \times P$.

After obtaining the global feature $\mathbf{x}_g$, we directly use a learnable MLP Head (MH) to map it into the global embedding with the same dimension as the label embedding:

$$\mathbf{e}_g = \mathrm{MH}\left(\mathbf{x}_g\right). \tag{3}$$

Knowledge Distillation: Here we also transfer knowledge from the VLP image encoder via knowledge distillation [43]. Specifically, we denote the frozen VLP image encoder as $\Phi_{\mathrm{I}}^{\mathrm{VLP}}(\cdot)$, and input the image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ into the VLP image encoder, and output:

$$\mathbf{e}_p = \Phi_{\mathrm{I}}^{\mathrm{VLP}}\left(\mathbf{x}\right). \tag{4}$$

To align $\mathbf{e}_g$ and the output $\mathbf{e}_p$ of the VLP image encoder, we apply $\mathcal{L}_1$ loss to minimize the distance between them, and the distillation process can be formulated as:

$$\mathcal{L}_{kd} = \mathbb{E}\left[\left\|\mathbf{e}_p - \mathbf{e}_g\right\|_1\right]. \tag{5}$$

Different Hierarchical Region Clues: Vanilla ViT divides the input image into very small patches of size $P \times P$, potentially missing objects. In contrast, integrating different hierarchical region clues can enhance the perception and recognition for object categories with different sizes in an image. For example, in Fig. 1, the smaller object category '*tennis*' and the larger object category '*dog*' in the image are recognized at different region hierarchies. our HRCP merges these patches hierarchically across three levels ($l = 1,\ 2,\ 3$), enabling the identification of potential objects at various patch sizes $[P * (4l + 1)]^2$, and hence is more suitable for multi-label classification. In order to gather region clues at different hierarchies, we first recover the spatial relationships among all original region representation output by Eq. (2):

$$\mathbf{H}_0^r = \begin{pmatrix} \mathbf{R_1} & \cdots & \mathbf{R}_N \\ \vdots & \ddots & \vdots \\ \mathbf{R}_{N^2-N+1} & \cdots & \mathbf{R}_{N^2} \end{pmatrix}, \tag{6}$$

where each element in $\mathbf{H}_0^r$ represents a region of size $P \times P$.

With the well-established spatial correlation, we then construct a convolution-based hierarchical merge operation to gather different hierarchical region clues $\{\mathbf{H}_l^r\}_{l=1}^3$ as shown in Fig. 2. Specifically, given the region features $\mathbf{H}_l^r$ at the *l*-th hierarchy, we obtain the features at (*l*+1)-th hierarchy by the hierarchical merge operation:

$$\begin{aligned} \mathbf{H}_l^{r\prime} &= \mathrm{Conv}(\mathrm{ReLU}(\mathrm{Conv}(\mathbf{H}_l^r))), \\ \mathbf{H}_{l+1}^r &= \mathrm{LN}(\mathbf{H}_l^{r\prime} + \mathrm{Bilinear}(\mathbf{H}_l^r)), \quad l = 0, 1, 2, \end{aligned} \tag{7}$$

where $\mathrm{Conv}(\cdot)$ denotes a 3×3 convolutional layer with a stride of 1 and without padding.

Finally, we use the same MLP Head (MH) as in Eq. (3) to map the obtained different hierarchical region features $\{\mathbf{H}_l^r\}_{l=1}^3$ into a joint visual-semantic embedding space:

$$\mathbf{E}_l^r = \mathrm{MH}\left(\mathbf{H}_l^r\right),\ l = 1, 2, 3. \tag{8}$$

D. Hierarchical Score Aggregation

We utilize scene-level global embeddings e_g and different hierarchical region embeddings $\{\mathbf{E}_l^r\}_{l=1}^3$ obtained through the above-mentioned hierarchical clue gathering module for multi-label classification. Typically, the global prediction score can be formulated as:

$$\mathbf{S}_{global} = \mathrm{sim}\left(\mathbf{e}_g,\ \mathbf{T}_s\right), \tag{9}$$

where $\mathbf{T}_s = [\mathbf{t}_1;\ \mathbf{t}_2;\ \cdots;\ \mathbf{t}_K]$ is the label embeddings of seen classes generated using the VLP text encoder, K denotes the number of seen classes, and $\mathrm{sim}(\cdot,\cdot)$ denotes dot product similarity.

Region Score Aggregation: For the hierarchical region embeddings $\{\mathbf{E}_l^r\}_{l=1}^3$, we propose a novel region score aggregation (RSA) approach to replace the *top-k* mean pooling used in previous methods [12], [20]. The motivation is to treat multi-label classification of an image as the single-label classification for each region within this image, and use softmax to obtain the probability that each region belongs to each class. Then, using these obtained probability scores, the original prediction scores of each region are weighted, with regions having higher relevance to the corresponding category being up-weighted and regions having lower relevance being down-weighted, thereby enhancing the discriminative power of the original prediction scores. Specifically, we first concatenate all hierarchical region embeddings together:

$$[\mathbf{e}_{r_1},\ \mathbf{e}_{r_2},\ \ldots,\ \mathbf{e}_{r_J}] = \mathrm{concat}\left(\mathbf{E}_1^r;\ \mathbf{E}_2^r;\ \mathbf{E}_3^r\right) \tag{10}$$

where $J = 140$ (i.e., $10^2 + 6^2 + 2^2$) denotes the total number of hierarchical region embeddings. Then, we compute the dot product similarity between each region embedding $\mathbf{e}_{r_j} \in \{\mathbf{e}_{r_1},\ \mathbf{e}_{r_2},\ \cdots,\ \mathbf{e}_{r_J}\}$ and the label embeddings $\mathbf{T}_s$ to obtain the corresponding region prediction score:

$$\mathbf{S}^j = \mathrm{sim}\left(\mathbf{e}_{r_j},\ \mathbf{T}_s\right). \tag{11}$$

Next, we apply the softmax function to each region score $\mathbf{S}^j \in \mathbb{R}^{1\times K}$ to obtain the probability that this region belongs to each seen category:

$$(P_{j,1}, P_{j,2}, \ldots, P_{j,K}) = \mathrm{softmax}\left(\mathbf{S}^j\right). \tag{12}$$

Finally, as shown in Fig. 2, we weight the initial region prediction scores based on the probability of each region belonging to each class to enhance discriminability, and then aggregate all weighted region prediction scores as the final region-based prediction:

$$\mathbf{S}_{region} = \sum_{j=1}^{J}\left[(P_{j,1},\ P_{j,2},\ \cdots,\ P_{j,K}) * \mathbf{S}^j\right], \tag{13}$$

where $*$ denotes element-wise multiplication. The prediction score of an image can be formulated as:

$$\bar{\mathbf{S}} = \lambda\mathbf{S}_{region} + (1 - \lambda)\mathbf{S}_{global}, \tag{14}$$

where λ is the hyper-parameter the controls the weights of $\mathbf{S}_{region}$ and $\mathbf{S}_{global}$. The network is trained with the ranking

loss on predicted scores, as follows:

$$\mathcal{L}_{rank} = \sum_{i} \sum_{p \in y_i, n \notin y_i} \max\left(\bar{\mathbf{S}}_i^n - \bar{\mathbf{S}}_i^p + 1,\ 0\right), \tag{15}$$

where $y_i \subseteq \mathcal{Y}^S$ are the seen class labels present in a multi-label image i, and $\bar{\mathbf{S}}_i^n$ and $\bar{\mathbf{S}}_i^p$ denote the scores of negative and positive labels, respectively.

Region Selection Strategy: Additionally, we design a novel region selection strategy (RSS) in the testing phase. Specifically, the j-th region prediction score $\mathbf{S}^j$ obtained by Eq. (11) can be written in detail as:

$$\mathbf{S}^j = \left(s_{j,1},\ s_{j,2},\ \cdots,\ s_{j,\hat{K}}\right), \tag{16}$$

where $s_{j,k}$ represents the dot product similarity between the j-th region embedding $\mathbf{e}_{r_j}$ and the k-th class label embedding $\mathbf{t}_k$, $\hat{K}$ denotes the number of test classes. We compute the sum of the dot product similarities of each region to the label embeddings of all test classes, and rank them:

$$\Gamma\left(\sum_{k=1}^{\hat{K}} s_{1,k},\ \sum_{k=1}^{\hat{K}} s_{2,k},\ \cdots,\ \sum_{k=1}^{\hat{K}} s_{J,k}\right), \tag{17}$$

where $\Gamma(\cdot,\ \cdots,\ \cdot)$ denotes a sort operation at descending order. We believe that low-ranked regions mean that they are not relevant to all test categories, belonging to classification-independent noise or background regions in an image. For all J (i.e., 140) hierarchical region embeddings, we remove the lowest-ranked 32 and use the remaining 108. With this strategy, we further improve the multi-label classification performance. It is worth noting that we only use this selection strategy during the testing phase, so no additional training is required.

E. Hybrid Prompt Learning

Since the predictions of the proposed HRCP rely on the dot product similarity between image embeddings and class label embeddings, and the class label embeddings are acquired by inputting the corresponding prompts (e.g., "A photo of a {class name}.") into the text encoder of the VLP model, the prompts used have a significant impact on the model's performance. This section explores how to obtain more suitable prompts for the OV-MLC task.

On the one hand, with carefully designed text prompts, VLP models show impressive generalization capabilities when transferred to downstream vision tasks [44], [45], but this requires specialized domain knowledge and careful text tuning. On the other hand, to avoid laborious prompt engineering, Context Optimization (CoOp) [46] has recently been proposed, which utilizes task-specific training data to learn continuous prompts to replace manually designed prompts. Despite improved performance on downstream tasks, some recent studies [47], [48] report that CoOp is prone to overfitting to the seen classes observed during training, resulting in poor performance on unseen classes. This indicates that CoOp impairs the generalization ability of the VLP model to unseen classes and out-of-distribution data.

Furthermore, since the text encoder of the VLP model is pre-trained based on global image-text embedding alignment, it cannot be directly applied to multi-label classification tasks. In order to make the VLP model better adapted to OV-MLC task, MKT [20] fine-tunes the token embedding layer of CLIP's text encoder to obtain better prompts. However, fine-tuning not only requires significant computational resources and memory consumption, but more importantly, it also poses the risk of forgetting the general visual-language knowledge in pre-trained CLIP, thus compromising the strong representation space learned on large-scale image-text pairs. Moreover, fine-tuning is more prone to overfitting on seen classes, especially when the dataset contains only a small number of seen class labels.

To address the aforementioned issues, we propose to combine learnable prompts with simple fixed prompts to make the generated label embeddings better match the corresponding visual embeddings, while preserving the generalization ability to unseen classes. Specifically, we first construct a learnable prompt:

$$\mathbf{P}_l = [\mathbf{v}_1\ \mathbf{v}_2\ \cdots\ \mathbf{v}_M], \tag{18}$$

where $\mathbf{v}_i$ represents the i-th learnable vector. In this work, we set $M = 8$, and initialize $\mathbf{P}_l$ with the embedding of "A photo for multi-label classification,".

Then, we employ a simple fixed prompt:

$$\mathbf{P}_m = [\text{there is a } \{\text{class name}\}.]. \tag{19}$$

Finally, the concatenation $\mathbf{P} = [\mathbf{P}_l,\ \mathbf{P}_m]$ of learnable prompt $\mathbf{P}_l$ and fixed manual prompt $\mathbf{P}_m$ is fed into the VLP text encoder and outputs the corresponding class label embedding $\mathbf{t}_k$. During prompt learning, the text encoder is completely frozen and simply optimized with the ranking loss for lightweight learnable prompt vectors. Experiments show that this hybrid prompt learning approach is beneficial for learning optimal prompt for the OV-MLC task.

F. Alternate Training

During training, the VLP image encoder and text encoder are kept frozen. We propose a simple yet effective alternative training strategy for the optimization of HCG and HPL modules (HSA module introduces no additional parameters and therefore does not require optimization). Specifically, in the HCG optimization stage, we fix learnable prompt vectors to generate label embeddings for seen classes, and use distillation loss and ranking loss for training:

$$\mathcal{L}_{image} = \mathcal{L}_{kd} + \mathcal{L}_{rank}. \tag{20}$$

During the prompt learning stage, we freeze the HCG, and use the ranking loss for training:

$$\mathcal{L}_{text} = \mathcal{L}_{rank}. \tag{21}$$

With this alternative training strategy, the image modality (i.e., HCG) and text modality (i.e., HPL) can mutually promote and improve together.

IV. Experiments

In this section, we first introduce the experimental setup, including datasets, evaluation metrics, and implementation details. After that, we compare the proposed HRCP with other state-of-the-art methods. Additionally, we conduct extensive ablation experiments to evaluate the gains brought by different components in HRCP, and provide visualization analysis to demonstrate the effectiveness of HRCP.

A. Experimental Setup

1) Datasets: We evaluate the proposed HRCP on NUS-WIDE [49] and Open Images [50] datasets. NUS-WIDE consists of 161,789 training images and 107,859 test images, each with 925 labels extracted from Flicker user tags and 81 human-annotated labels. Following previous work [20], 925 and 81 labels are used as seen and unseen classes, respectively. Open Images (v4) is a large-scale dataset containing nearly 9 million training images and 125,456 testing images. Consistent with previous work [20], 7186 labels with more than 100 training images are selected as seen classes, and the 400 test set labels that appear least frequently in the training data are selected as the unseen classes.

2) Evaluation Metrics: Following previous works [20], we use F1 score at *top-K* predictions [33], [51] and mean Average Precision (mAP) [52] as evaluation metrics. Specifically, the F1 score at *top-K* predictions measures how accurately the model ranks the labels in each image. It focuses on the prediction accuracy of the *top-K* most prominent/significant categories in a multi-label image. Meanwhile, mAP captures the model's ability to correctly rank images for each label, i.e., the accuracy of retrieving relevant images based on labels. Furthermore, considering that the F1 score is derived from the harmonic mean of precision (P) and recall (R), i.e., $F1 = 2 * P * R/(P + R)$, we also report precision and recall in our experiments for a comprehensive evaluation.

3) Implementation Details: For a fair comparison with MKT [20], we choose pre-trained and frozen CLIP based on ViT-B/16 as our VLP model. The ViT in the HCG module is also a ViT-B/16, the resolution of input images is 224×224, and the resulting number of patches is $N^2 = 196$. The MLP Head is a linear projection layer. We adopt the AdamW optimizer [53] with a weight decay of 0.05 and a base learning rate of 0.001/0.0001 for NUS-WIDE/Open Images. For NUS-WIDE, we use batch sizes of 64 and 16 for the HCG optimization stage and prompt learning optimization stage, respectively. We initially train the HCG for 10 epochs, and subsequently optimize the learnable prompt for 1 epoch. This alternating process is repeated 5 times on NUS-WIDE, resulting in a cumulative training of 50 epochs for the HCG and 4 epochs for the learnable prompt. For Open Images, we use a batch size of 128 for all train stages, and the alternate training process is as follows: we first train the HCG for 3 epochs, then optimize the learnable prompt for 1 epoch, and finally continue training the HCG for an additional 2 epochs. We set $\lambda = 0.25$ and $\lambda = 0.5$ for the NUS-WIDE and Open Images, respectively. HRCP is implemented with PyTorch, and all the experiments are performed on an NVIDIA GeForce RTX 3090 GPU.

B. State-of-The-Art Comparison

Tabel I presents the state-of-the-art performance comparison on Zero-shot learning (ZSL) and Generalized zero-shot learning (GZSL) tasks. Among them, LESA [10], ZS-SDL [11], BiAM [12], Gen-MLZSL [35], and $(ML)^2$P-Encoder [13] are the previous state-of-the-art ZS-MLC methods. These methods utilize language models pre-trained on large-scale corpora to transfer knowledge from seen classes to unseen classes, only exploring the knowledge transfer in the textual modality. On the other hand, OVML-VLP [21], CLIP-FT [20], and MKT [20] are OV-MLC methods based on VLP models. OVML-VLP [21] extracts multiple class-specific image features from the original CLIP using an image-text attention module, thereby enabling generalization to unseen classes. CLIP-FT is an OV-MLC baseline proposed by [20] that fine-tunes pre-trained CLIP on the seen classes according to the ranking loss. MKT [20] utilizes knowledge distillation to transfer knowledge from the VLP image encoder, and employs prompt tuning to enable the label embeddings generated by the VLP text encoder to better support OV-MLC tasks. OVML-VLP and MKT are currently the state-of-the-art methods for OV-MLC. They achieve better performance than ZS-MLC methods, which illustrates that exploring the multi-modal knowledge of image-text pairs from VLP models can more effectively identify unseen labels.

It can also be observed that, for NUS-WIDE dataset, HRCP achieves superior mAP scores on both ZSL and GZSL tasks compared to the current state-of-the-art techniques, with absolute gains up to 3.2% (ZSL) and 4.4% (GZSL). Regarding the F1 scores at $K \in \{3, 5\}$, HRCP achieves the second-best and best results on ZSL and GZSL tasks, respectively. On Open Images, for the ZSL task, HRCP achieves state-of-the-art performance across all metrics, especially mAP, with an absolute gain of up to 3.5% compared to MKT, raising the score from 68.1% to 71.6% . For the GZSL task, HRCP also achieves the best F1 score at $K \in \{10, 20\}$ and competitive (second best) mAP. The comprehensive improvement in performance on both datasets demonstrates the superiority of the proposed method. It should be noted that due to the complexity and challenges of Open Images, in some metrics, even MKT only achieved a slight improvement over the fine-tuned CLIP (i.e., CLIP-FT), while our HRCP achieved a larger improvement relative to MKT. For example, the F1 scores at $K = 10$ are as follows: CLIP-FT (19.1%) vs MKT (19.7%) vs HRCP (20.6%). Experimental results demonstrate the effectiveness of HRCP, which can better recognize multiple object categories at different sizes in a multi-label image by gathering various hierarchical region clues.

C. mAP Improvement Comparison

For a more comprehensive evaluation, we conduct Average Precision (AP) comparisons for each class with the previous state-of-the-art method MKT [20] on the NUS-WIDE dataset. Across all 81 unseen classes, our HRCP surpasses MKT in AP

TABLE I

STATE-OF-THE-ART COMPARISON ON THE NUS-WIDE AND OPEN IMAGES DATASETS. WE REPORT THE RESULTS IN TERMS OF MAP, AS WELL AS PRECISION (P), RECALL (R), AND F1 SCORE AT $K \in \{3, 5\}$ FOR NUS-WIDE AND $K \in \{10, 20\}$ FOR OPEN IMAGES. THE BEST AND SECOND BEST RESULTS ARE MARKED IN **BOLD** AND RED, RESPECTIVELY. SINCE BIAM [12] REPORTS WEIGHTED-MAP ON OPEN IMAGES, FOR A FAIR COMPARISON WITH OTHER METHODS, WE RE-IMPLEMENTED BIAM'S MAP USING THE AUTHOR'S OFFICIAL CODE, THE RESULTS ARE HIGHLIGHTED WITH SYMBOL $*$.

Methods	Setting	Task	NUS-WIDE (#seen / #unseen = 925 / 81) K=3 P	K=3 R	K=3 F1	K=5 P	K=5 R	K=5 F1	mAP	Open Images (#seen / #unseen = 7186 / 400) K=10 P	K=10 R	K=10 F1	K=20 P	K=20 R	K=20 F1	mAP
LESA (M=10) CVPR'2020 [10]	ZS	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
ZS-SDL ICCV'2021 [11]	ZS	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
BiAM ICCV'2021 [12]	ZS	ZSL	-	-	33.1	-	-	30.7	26.3	-	-	8.3	-	-	5.5	62.2*
		GZSL	-	-	16.1	-	-	19.0	9.3	-	-	19.1	-	-	15.9	67.1*
Gen-MLZSL TPAMI'2023 [35]	ZS	ZSL	26.6	42.8	32.8	20.1	53.6	29.3	25.7	1.3	42.4	2.5	1.1	52.1	2.2	43.0
		GZSL	30.9	13.6	18.9	26.0	19.1	22.0	8.9	33.6	38.9	36.1	22.8	52.8	31.9	75.5
$(ML)^2P$-Encoder CVPR'2023 [13]	ZS	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
OVML-VLP ICME'2023 [21]	OV	ZSL	**36.3**	44.7	**40.1**	**27.9**	57.2	**37.5**	42.6	9.6	74.4	16.9	5.6	87.5	10.6	68.4
		GZSL	32.9	12.6	18.3	28.1	18.0	22.0	14.3	33.9	39.2	36.4	23.2	53.5	32.3	77.8
CLIP-FT AAAI'2023 [20]	OV	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
MKT AAAI'2023 [20]	OV	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**
HRCP [ours]	OV	ZSL	31.5	**50.3**	38.7	24.1	**64.2**	35.1	**45.8**	**11.6**	**90.7**	**20.6**	**6.2**	**96.5**	**11.6**	**71.6**
		GZSL	**40.5**	**17.9**	**24.8**	**33.9**	**24.9**	**28.7**	**22.7**	**40.1**	**46.3**	**43.0**	**26.6**	**61.4**	**37.1**	79.9

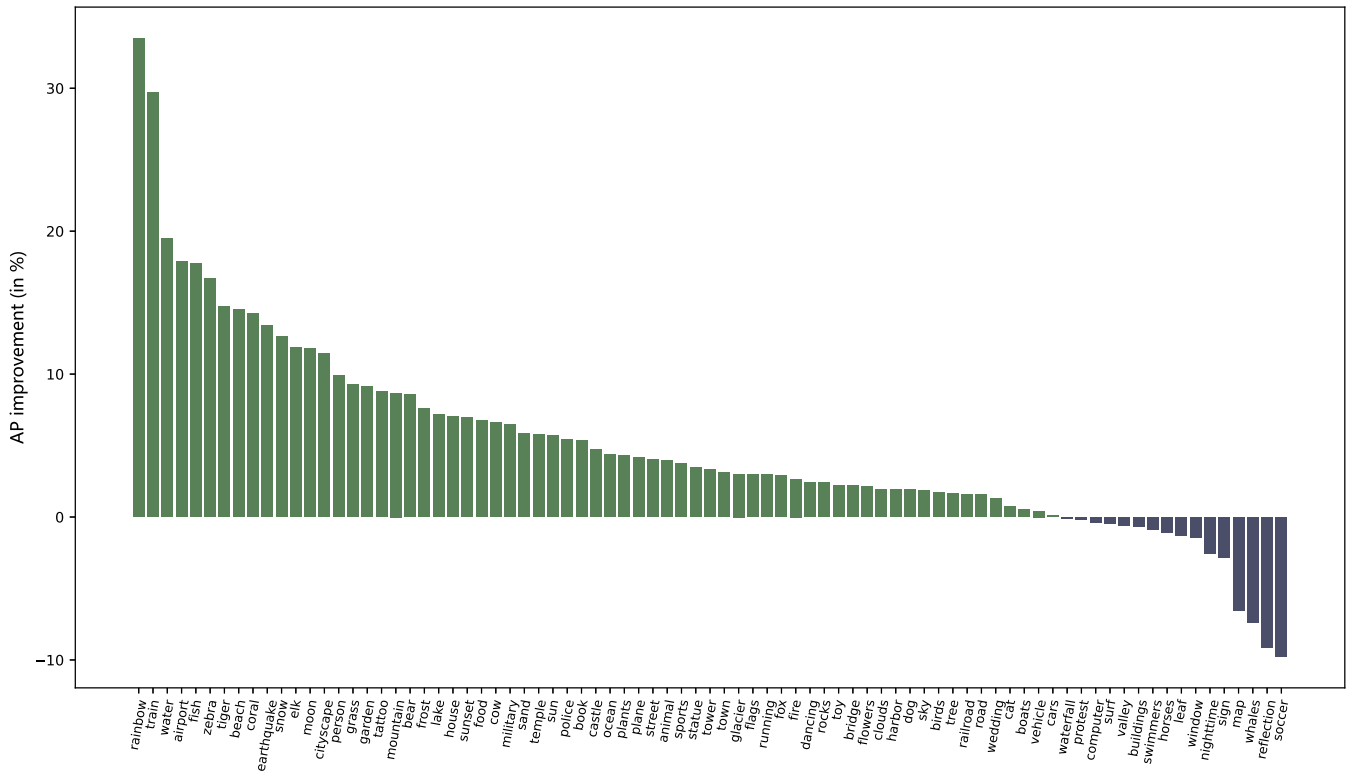


Fig. 3. Comparison of mAP improvement between our HRCP and MKT [20] on NUS-WIDE. In all 81 unseen classes, our HRCP outperforms MKT in terms of AP on 65 classes.

TABLE II
ABLATION STUDIES FOR HCG AND HSA ON NUS-WIDE. THE BEST RESULTS ARE MARKED IN **BOLD**.

HCG	RSA	RSS	K=3			K=5			mAP
			P	R	F1	P	R	F1	
✗	✗	✗	27.8	44.5	34.2	21.4	56.9	31.1	43.9
✓	✗	✗	29.8	47.6	36.7	23.1	61.6	33.6	44.1
✓	✓	✗	31.3	50.1	38.6	24.0	64.0	34.9	45.3
✓	✓	✓	**31.5**	**50.3**	**38.7**	**24.1**	**64.2**	**35.1**	**45.8**

TABLE III
COMPARISON OF DIFFERENT PROMPTS ON NUS-WIDE. THE BEST RESULTS ARE MARKED IN **BOLD**.

Prompt	Params	Train/epoch	F1(K=3)	F1(K=5)	mAP
Fixed Prompt	0	0 min	32.0	28.8	**45.9**
Tuning Embedding	25.3M	151 min	37.3	33.4	44.4
Learnable Prompt	0.004M	84 min	37.9	34.0	45.1
Learnable + Fixed	0.004M	84 min	**38.7**	**35.1**	45.8

for 65 classes. As illustrated in Fig. 3, HRCP demonstrates significant improvement (over 30%) for certain unseen labels, such as '*rainbow*', while exhibiting a relatively minor negative impact (less than 10%) on labels like '*soccer*'. This further verifies the superiority of our method. For the few categories (e.g., '*soccer*') where HRCP's AP is lower than MKT, we analyze that this could be attributed to HRCP's strategy of balancing attention across multiple-sized objects in an image by gathering different hierarchical region clues and scene-level global clues, whereas MKT focuses more on global objects and small objects of size $P \times P$ (i.e., 16×16 in ViT-B/16). As a result, HRCP's attention on certain specific-sized object categories may not be as concentrated as MKT's. Nevertheless, HRCP's overall advantages remain significant.

D. Ablation Studies

1) Ablation Study for HCG and HSA: To evaluate the contribution of each module in HRCP, we conduct an ablation study as shown in Tabel II. HCG represents the proposed hierarchical clue gathering module, while RSA and RSS respectively denote the region score aggregation and region selection strategy in the proposed hierarchical score aggregation (HSA) module. As can be seen from Tabel II, employing HCG to gather hierarchical region embeddings (the second row) significantly improves the classification accuracy compared to solely using the original region embeddings output by ViT (the first row). This highlights the beneficial role of hierarchical region clues in recognize multiple categories of different sizes in an image. Subsequently, replacing *top-k* mean pooling with RSA to aggregate hierarchical region prediction scores (third row) further enhances performance. This is because *top-k* mean pooling directly averages the *top-k* predictions for each category over all image regions, without fully utilizing all predictions of each class across all regions, and a simple average will lead to relatively smooth results, which reduces the discriminability. While the proposed RSA more effectively utilizes the predictions of each regions for each category. Finally, the introduction of RSS (fourth row) yields the best results, indicating its effectiveness in removes noise or background areas that are not relevant to classification. The ablation study demonstrates that different components promote each other and work together to better unleash the potential of hierarchical region clues in images.

2) Comparison of Different Prompts: To validate the effectiveness of the adopted hybrid prompt learning strategy, we compare it with other prompt learning methods on the ZSL task, and the results are presented in Tabel III. It can be observed that **Fixed Prompt** (i.e., fixing both $\mathbf{P}_l$ and $\mathbf{P}_m$) has strong generalization, which is reflected in the fact that the model can accurately retrieve relevant images according to the given unseen labels (i.e., high mAP). Since mAP aims to capture the ranking accuracy of all images for each label, and is typically used to evaluate the overall performance of the model on all categories, it is more suitable as a metric to evaluate the generalization ability of the model. However, since the VLP model is learned from global image-text alignments and cannot be directly applied to multi-label classification tasks, HRCP utilizing only fixed prompt cannot predict and rank multiple labels in an image well (i.e., low F1 score). The F1 score only focuses on the prediction accuracy of the *top-K* most prominent/salient categories in each image, and is therefore often used to evaluate the model's ability to identify the most important categories in a specific task. However, the F1 score does not take into account the overall ranking accuracy of all labels for each image. **Tuning Embedding** (i.e., fine-tuning the token embedding layer of the text encoder as in MKT [20]) and **Learnable Prompt** (i.e., making $\mathbf{P}_l$ and $\mathbf{P}_m$ all learnable) can better adapt to the OV-MLC task by training on the seen class data, which is reflected in the improvement of the F1 score; but they are prone to overfitting on seen classes, so the mAP on unseen classes drops. Tuning Embedding has a more serious overfitting problem than Learnable Prompt, because fine-tuning VLP's text encoder impairs its strong representation and generalization capabilities. On the other hand, since fine-tuning the token embedding layer (25.3M) introduces more trainable parameters than Learnable Prompt (0.004M), thus requiring a longer training time. For example, training learnable prompts on NUS-WIDE only takes 84 minutes per epoch, while fine-tuning the token embedding layer takes 151 minutes. This further illustrates the superiority of learnable prompts over fine-tuning the token embedding layer. Finally, **Learnable + Fixed** (i.e., our hybrid prompt learning method, where $\mathbf{P}_l$ is learnable and $\mathbf{P}_m$ is fixed) achieves the best performance by combining the advantages of fixed and learnable prompt.

3) The Impact of Knowledge Distillation: The purpose of knowledge distillation is to transfer rich multi-modal knowledge from pre-trained VLP models, making it a popular solution for OV-based tasks [20], [25], [26], [44]. To clarify the role of knowledge distillation in the proposed HRCP, we removed the knowledge distillation loss and compared the performance with the full model that includes it. The experimental results, as shown in Table IV, indicate that the overall performance of the model significantly declines after removing the knowledge distillation

TABLE IV
THE IMPACT OF KNOWLEDGE DISTILLATION ON NUS-WIDE. THE BEST RESULTS ARE MARKED IN **BOLD**.

Models	Task	K=3			K=5			mAP
		P	R	F1	P	R	F1	
HRCP (w/o $\mathcal{L}_{kd}$)	ZSL	22.3	35.6	27.4	16.6	44.2	24.1	33.5
	GZSL	35.6	15.7	21.8	29.4	21.6	24.9	19.0
HRCP	ZSL	**31.5**	**50.3**	**38.7**	**24.1**	**64.2**	**35.1**	**45.8**
	GZSL	**40.5**	**17.9**	**24.8**	**33.9**	**24.9**	**28.7**	**22.7**

TABLE V
COMPARISON WITH ORIGINAL PRE-TRAINED CLIP ON NUS-WIDE. THE BEST RESULTS ARE MARKED IN **BOLD**.

Models	Task	K=3			K=5			mAP
		P	R	F1	P	R	F1	
Pre-CLIP	ZSL	26.7	42.7	32.8	19.1	50.8	27.7	27.6
	GZSL	10.2	4.5	6.3	8.8	6.5	7.5	10.5
HRCP	ZSL	**31.5**	**50.3**	**38.7**	**24.1**	**64.2**	**35.1**	**45.8**
	GZSL	**40.5**	**17.9**	**24.8**	**33.9**	**24.9**	**28.7**	**22.7**

module. This demonstrates the necessity and effectiveness of transferring knowledge from VLP image encoders via knowledge distillation, as it helps the model better extract features, improve classification accuracy, and enhance robustness, playing a crucial role in strengthening the model's generalization ability. On the other hand, without using knowledge distillation to transfer multi-modal knowledge from VLP models, our model degenerates into a traditional zero-shot multi-label classification model. However, its performance still surpasses the current state-of-the-art ZS-MLC methods, further validating the effectiveness and advantages of our proposed HRCP model.

4) Comparison With Original Pre-Trained CLIP: To better highlight the contributions of this paper, we performed experimental evaluations of the original pre-trained CLIP model and compared its performance with the proposed HRCP. The results in Table V demonstrate that the performance of the original pre-trained CLIP on ZSL and GZSL tasks is significantly lower than that of HRCP and even inferior to some traditional ZS-MLC methods listed in Table I. For example, the original pre-trained CLIP achieves an mAP of only 27.6% (10.5%) on the ZSL (GZSL) task, whereas HRCP achieves an mAP of 45.8% (22.7%). This discrepancy arises because CLIP is trained to align global image embeddings with text embeddings, focusing solely on global information within an image. As a result, it performs better on scene-level single-label classification tasks. However, the OV-MLC task requires the model to accurately identify multiple object categories of different sizes in multi-label scenarios. CLIP fails to effectively capture local information in multi-label images and ignores labels associated with smaller regions, limiting its performance on multi-label tasks. Additionally, its text prompts are not optimized for multi-label classification tasks, further restricting its generalization performance. In contrast, the proposed method addresses these shortcomings of CLIP in the OV-MLC task by introducing the hierarchical clue gathering (HCG) module, the hierarchical score aggregation (HSA) module, and the hybrid prompt learning (HPL) module. Experimental results show that HRCP significantly improves the performance of OV-MLC, validating the effectiveness and innovation of the proposed approach.

E. Hyper-Parameter Sensitivity

1) The Effect of Different λ: λ is the hyper-parameter in Eq. (14) that controls the weight of the global and regional prediction scores. Fig. 4(a) illustrates the impact of changes in λ on the prediction accuracy of the model on the NUS-WIDE dataset. The larger λ indicates that the weight of region prediction is larger, $\lambda = 0$ and $\lambda = 1$ indicate that only global and regional prediction are used, respectively. It is evident that excessively small or large values of λ lead to suboptimal performance, which shows that scene-level global clues and hierarchical region clues both are very important, and they work together to achieve the recognition of multiple categories with different sizes in an image. Notably, the optimal performance on NUS-WIDE is achieved when $\lambda = 0.25$. Additionally, for Open Images, the best results are attained at $\lambda = 0.5$, which may be because Open Images contains more categories than NUS WIDE, so more attention needs to be paid to the region-based prediction.

2) Select Different Number of Regions With RSS: In Tabel II, we demonstrate the effectiveness of the region selection strategy (RSS) through ablation experiments on model components for the ZSL task on NUS-WIDE. Fig. 4(b) further presents the results of using RSS to select different numbers of hierarchical region embeddings for the more challenging GZSL task on the larger Open Images dataset. It can be seen that the prediction accuracy first improves as the number of selected regions decreases, because our RSS removes noise or background regions in the image that are irrelevant to classification. Subsequently, as the number of regions continues to decrease, the prediction accuracy starts to drop due to the over-removal of class-related regions. For both NUS-WIDE and Open Images, we utilize RSS to select 108 regional features for prediction without additional adjustments for specific datasets.

3) The Effect of Different M: M is a hyper-parameter in Eq. (18) that determines the number of vectors in the learnable prompt, directly influencing the model's ability to capture complex context information in multi-label tasks. This section studies the impact of varying the hyper-parameter M on the model's performance on the NUS-WIDE dataset, with the experimental results shown in Fig. 4(c). As observed, the model achieves optimal performance when $M = 8$, while overly small or large values of M adversely affect performance. This is because a smaller M (e.g., 4) may limit the capability of prompt learning, thereby affecting model performance. In contrast, a larger M (e.g., 16) allows the model to learn richer context information but may also increase the risk of overfitting. Therefore, M is set to 8 in this paper to achieve a good trade-off between learning capability and overfitting risk, resulting in optimal performance.

F. Number of Parameters and Inference Time

Table VI presents the comparison of the number of learnable parameters and performance of different methods. Among them,

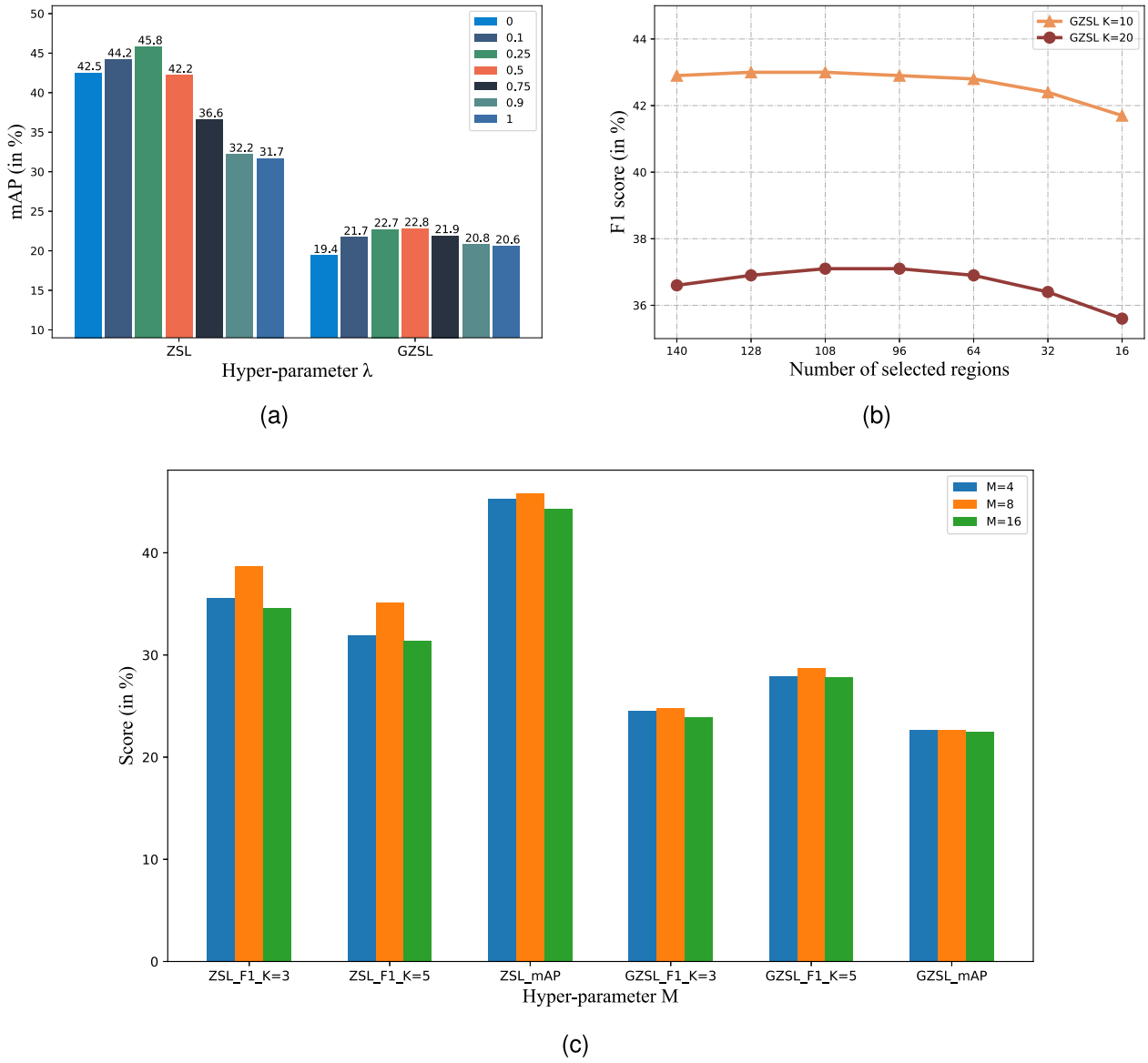


Fig. 4. Hyper-parameter sensitivity. (a) The impact of different λ. (b) Select different number of regions with RSS. (c) The impact of different M.

TABLE VI
COMPARISON OF THE NUMBER OF TRAINABLE PARAMETERS AND PERFORMANCE ON NUS-WIDE

Methods	Parameters			F1		mAP
	Image encoder	Text encoder	ALL	K=3	K=5	
LESA (M=10)	0.45M	/	0.45M	31.6	28.7	19.4
ZS-SDL	33.6M	/	33.6M	30.5	27.8	25.9
BiAM	3.8M	/	3.8M	32.7	29.8	25.9
Gen-MLZSL	216.0M	/	216.0M	32.8	29.3	25.7
MKT	15.2M	25.3M	40.5M	34.1	31.1	37.6
HRCP [ours]	18.3M	0.004M	18.3M	**38.7**	**35.1**	**45.8**

ZS-SDL [11] and BiAM [12] are traditional ZS-MLC methods that do not utilize the VLP model, while MKT [20] and the proposed HRCP are VLP-based OV-MLC methods. Both MKT and the proposed HRCP utilize pre-trained CLIP and ViT-B/16, where the image encoder of CLIP is completely frozen. Additionally, MKT freezes the first 10 layers of ViT-B/16, while HRCP freezes the first 11 layers. Furthermore, MKT fine-tunes the token embedding layer of CLIP text encoder, whereas HRCP learns lightweight learnable prompt. Overall, the number of learnable parameters of HRCP (18.3 M) is much smaller than MKT (40.5 M). However, HRCP achieves higher mAP and F1 scores compared to MKT, indicating that performance improvement comes from methodological innovation rather than an increase in the number of parameters. Moreover, HRCP also has smaller learnable overhead and better performance compared with ZS-SDL and Gen-MLZSL that do not utilize VLP models. This further validates the effectiveness of HRCP.

In addition, we also compared the time required for several methods to infer the ZSL and GZSL prediction scores of a single image: BiAM (4.7 ms) vs MKT (10.6 ms) vs HRCP (12.5 ms). The experiments are conducted on the same NVIDIA GeForce

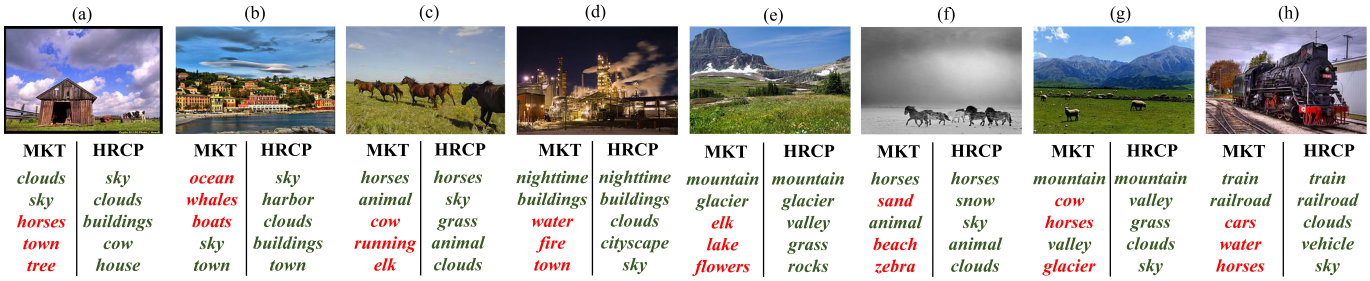


Fig. 5. Prediction results on several unseen exemplar images from the NUS-WIDE test set. The *top-5* predictions per image for both methods are shown as *true positives* and *false positives*. Best viewed in color.

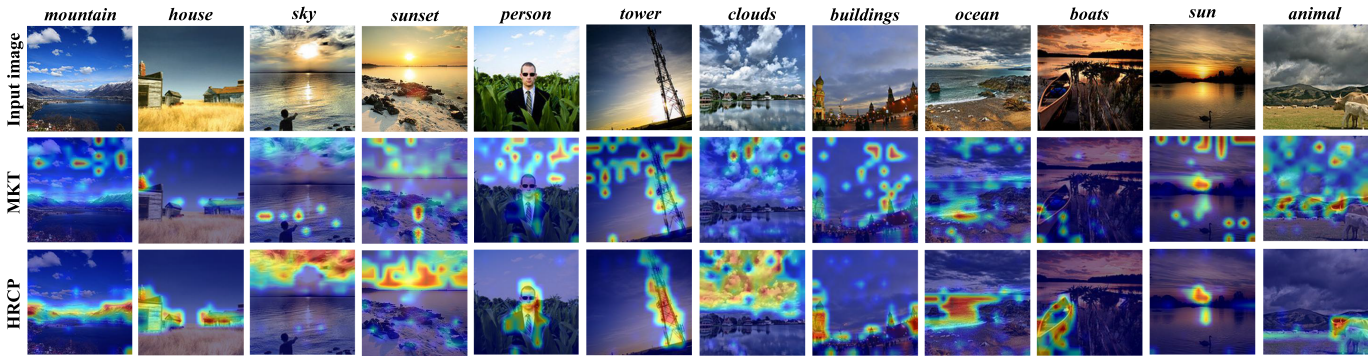


Fig. 6. Comparison of Grad-CAM visualization on specific unseen classes. For each image, the class-specific map of the ground truth unseen class are shown, with the corresponding label on top.

RTX 3090 GPU. It can be observed that due to the introduction of more hierarchical region clues, HRCP slightly decreases the inference speed for a single image compared to BiAM and MKT. However, considering the significant increase in performance, this reduction is negligible. Moreover, this speed (i.e., 12.5 ms for an image) is sufficient to ensure the real-time application of the model. Note that BiAM starts inference from the extracted VGG features, while MKT and HRCP infer directly from the original image. This is also a reason why BiAM achieves faster inference speed. Finally, we also compared the total time required by BiAM, MKT, and HRCP to process the entire NUS-WIDE test dataset consisting of 107,859 images under the same batch size (i.e., 471) setting: BiAM (277 s) vs MKT (121 s) vs HRCP (167 s). It can be found that although BiAM has a slight advantage in inference speed for a single image, MKT and HRCP are faster when handling the entire dataset. Additionally, despite the proposed HRCP introduces multiple hierarchical region clues, it is only 46 seconds slower than MKT when processing more than 100,000 images from the entire NUS-WIDE test dataset, which further confirms the efficiency of HRCP in terms of inference speed.

In summary, by freezing the parameters of the VLP model and introducing a lightweight hybrid prompt learning strategy, HRCP significantly reduces the computational resource requirements, ensuring better scalability for training and inference on larger-scale datasets. Additionally, the proposed hierarchical clue gathering module and hybrid prompt learning strategy are not dependent on a specific number of categories, making them adaptable to large-scale, multi-category scenarios. When handling large-scale datasets, we can further explore techniques such as knowledge distillation or pruning to generate smaller models, thereby improving the practicality and efficiency of the model while maintaining high performance.

G. Qualitative Results

Fig. 5 presents the *top-5* unseen labels predicted by MKT [20] and our HRCP for some test images from NUS-WIDE. The tags in olive green appear in ground-truth annotations, and those in red are wrong tags. Compared with MKT, our method produces more accurate and diverse predictions. For example, in Fig. 5(a), MKT only recognizes the categories '*clouds*' and '*sky*', while our HRCP further makes accurate predictions for the categories '*cow*', '*house*', and '*buildings*' with different sizes. Similarly, in Fig. 5(c), MKT only predicts the classes '*horses* and '*animal*', while HRCP also recognizes the classes '*sky*', '*grass*', and '*clouds*'. Moreover, in Fig. 5(e), MKT only predicts the categories '*mountain* and '*glacier*', whereas HRCP also recognizes the categories '*valley*', '*grass*', and '*rocks*'. This demonstrates the effectiveness of HRCP, which can better recognize multiple object categories at different scales in an image by integrating scene-level global clues and different hierarchical region clues.

H. Grad-CAM Visualization

Fig. 6 shows the comparison of Grad-CAM visualization for specific unseen categories in example test images from

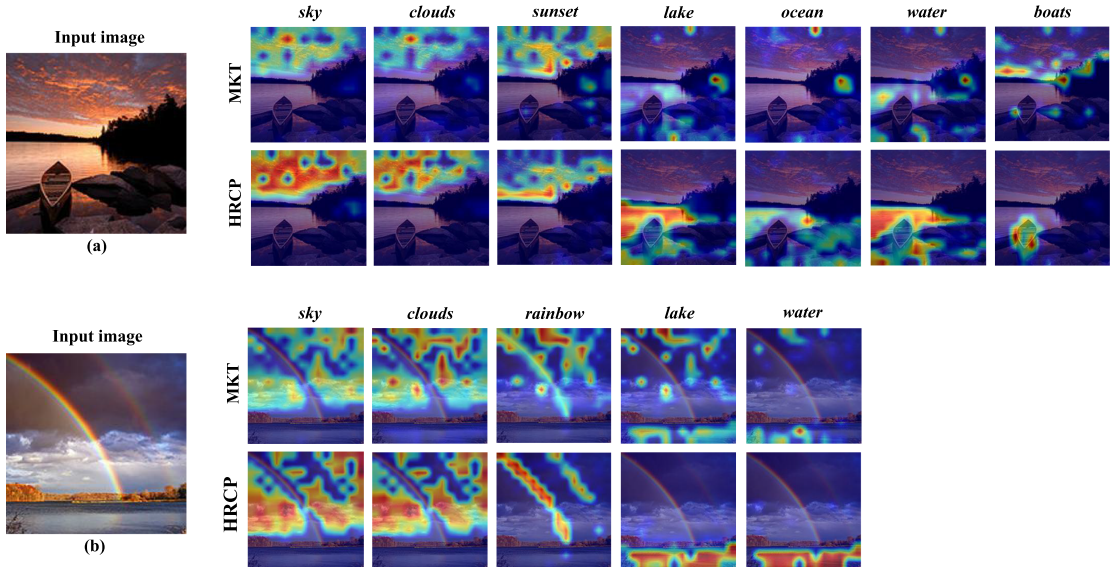


Fig. 7. Comparison of Grad-CAM visualization on specific images. For each image, the class-specific maps for all ground truth unseen classes are shown, with the corresponding labels on top.

NUS-WIDE between MKT [20] and our HRCP. Along with each example, we also present the class-specific mappings for each unseen class, with the corresponding labels on top. It can be observed that MKT generates dispersed attention, while HRCP more accurately captures the class-specific relevant region. This is because MKT only utilizes single hierarchical region clues, focuses on perceiving small regions of size $P \times P$, leading to weaker recognition capability for object categories at other sizes in the image. Fig. 7 further presents the Grad-CAM visualization for all ground-truth unseen labels present in the corresponding multi-label example test image from NUS-WIDE. It can be observed that compared with MKT, our HRCP can capture the relevant regions of almost all ground-truth classes in an image more precisely. This demonstrates that by fully unleashing the potential of hierarchical region clues, HRCP can better perceive multiple categories with different sizes in an image, thereby generate promising category-specific attention maps.

I. Further Discussion

The HRCP model proposed in this paper is primarily designed and evaluated for multi-label classification tasks in static images. However, due to its modular design, the core ideas of HRCP are highly flexible and generalizable, allowing for easy adaptation to other types of multimedia data. For example, by incorporating temporal feature modeling into the hierarchical clue gathering module, HRCP can be extended to open-vocabulary video multi-label classification tasks. Additionally, HRCP's multi-modal alignment capabilities (particularly in knowledge distillation and hybrid prompt learning) are also applicable to tasks that combine visual, audio, and textual information, such as audio-video synchronization or speech-driven image understanding tasks. Due to space limitations, we will explore these aspects as part of our future work to further investigate the potential of HRCP.

V. Conclusion

This paper addresses the open-vocabulary multi-label classification task by fully unleashing the potential of hierarchical region clues. First, we gather hierarchical region clues and scene-level global clues to facilitate the recognition of multiple object categories with different sizes in an multi-label image. Moreover, by using a novel hierarchical region score aggregation approach, we effectively utilize the prediction score of each region for each category, thus obtaining more discriminative result. Additionally, we propose a hybrid prompt learning method to generate label embeddings that are better adapted to the OV-MLC task. Finally, we also use a region selection strategy to remove noise or background regions that are irrelevant to the classification, thereby further improve prediction accuracy. Extensive experimental results and analyses validate the effectiveness of the proposed framework.

References

[1] F. Lyu, Q. Wu, F. Hu, Q. Wu, and M. Tan, "Attend and Imagine: Multi-label image classification with visual attention and recurrent neural networks," *IEEE Trans. Multimedia*, vol. 21, pp. 1971–1981, 2019.
[2] J. Xu et al., "Joint input and output space learning for multi-label image classification," *IEEE Trans. Multimedia*, vol. 23, pp. 1696–1707, 2021.
[3] F. Zhou, S. Huang, and Y. Xing, "Deep semantic dictionary learning for multi-label image classification," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, pp. 3572–3580.
[4] X. Deng, S. Feng, G. Lyu, T. Wang, and C. Lang, "Beyond word embeddings: Heterogeneous prior knowledge driven multi-label image classification," *IEEE Trans. Multimedia*, vol. 25, pp. 4013–4025, 2023.
[5] R. You et al., "Cross-modality attention with semantic graph embedding for multi-label classification," in *Proc. AAAI Conf. Atificial Intell.*, 2020, vol. 34, pp. 709–716.
[6] T. Ridnik et al., "Asymmetric loss for multi-label classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 82–91.
[7] J. Lanchantin, T. Wang, V. Ordonez, and Y. Qi, "General multi-label image classification with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16478–16488.

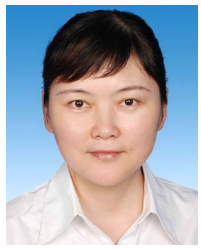
[8] Y. Wang, T. Zhang, C. Zhou, Z. Cui, and J. Yang, "Instance-aware deep graph learning for multi-label classification," *IEEE Trans. Multimedia*, vol. 25, pp. 90–99, 2023.
[9] W. Zhou, W. Jiang, D. Chen, H. Hu, and T. Su, "Mining semantic information with dual relation graph network for multi-label image classification," *IEEE Trans. Multimedia*, vol. 26, pp. 1143–1157, 2024.
[10] D. Huynh and E. Elhamifar, "A shared multi-attention framework for multi-label zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8776–8786.
[11] A. Ben-Cohen, N. Zamir, E. Ben-Baruch, I. Friedman, and L. Zelnik-Manor, "Semantic diversity learning for zero-shot multi-label classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 640–650.
[12] S. Narayan et al., "Discriminative region-based multi-label zero-shot learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 8731–8740.
[13] Z. Liu et al., "$(ML)^2$P-Encoder: On exploration of channel-class correlation for multi-label zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 23859–23868.
[14] Z. Ye, F. Hu, F. Lyu, L. Li, and K. Huang, "Disentangling semantic-to-visual confusion for zero-shot learning," *IEEE Trans. Multimedia*, vol. 24, pp. 2828–2840, 2022.
[15] X. Chen, J. Li, X. Lan, and N. Zheng, "Generalized zero-shot learning via multi-modal aggregated posterior aligning neural network," *IEEE Trans. Multimedia*, vol. 24, pp. 177–187, 2022.
[16] Y. Yang, X. Zhang, M. Yang, and C. Deng, "Adaptive bias-aware feature generation for generalized zero-shot learning," *IEEE Trans. Multimedia*, vol. 25, pp. 280–290, 2023.
[17] Y. Li, Z. Liu, L. Yao, and X. Chang, "Attribute-modulated generative meta learning for zero-shot learning," *IEEE Trans. Multimedia*, vol. 25, pp. 1600–1610, 2023.
[18] R. Gao et al., "Visual-semantic aligned bidirectional network for zero-shot learning," *IEEE Trans. Multimedia*, vol. 25, pp. 1649–1664, 2023.
[19] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.
[20] S. He et al., "Open-vocabulary multi-label classification via multi-modal knowledge transfer," in *Proc. AAAI Conf. Artif. Intell.*, 2023, vol. 37, pp. 808–816.
[21] S. D. Dao, D. Huynh, H. Zhao, D. Phung, and J. Cai, "Open-vocabulary multi-label image classification with pretrained vision-language model," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2023, pp. 2135–2140.
[22] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[23] S. Zhao et al., "Exploiting unlabeled data with vision and language models for object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 159–175.
[24] S. Wu, W. Zhang, S. Jin, W. Liu, and C. C. Loy, "Aligning bag of regions for open-vocabulary object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 15254–15264.
[25] L. Li et al., "Distilling DETR with visual-linguistic knowledge for open-vocabulary object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 6501–6510.
[26] L. Wang et al., "Object-aware distillation pyramid for open-vocabulary object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 11186–11196.
[27] C. Ma, Y. Jiang, X. Wen, Z. Yuan, and X. Qi, "CoDet: Co-occurrence guided region-word alignment for open-vocabulary object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, vol. 36, pp. 1–17.
[28] M. Xu et al., "A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 736–753.
[29] J. Mukhoti et al., "Open vocabulary semantic segmentation with patch aligned contrastive learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19413–19423.
[30] F. Liang et al., "Open-vocabulary semantic segmentation with mask-adapted clip," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 7061–7070.
[31] X. Xu, T. Xiong, Z. Ding, and Z. Tu, "MasQCLIP for open-vocabulary universal image segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 887–898.
[32] H. Luo, J. Bao, Y. Wu, X. He, and T. Li, "SegCLIP: Patch aggregation with learnable centers for open-vocabulary semantic segmentation," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 23033–23044.
[33] Y. Zhang, B. Gong, and M. Shah, "Fast zero-shot image tagging," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5985–5994.
[34] C.-W. Lee, W. Fang, C.-K. Yeh, and Y.-C. F. Wang, "Multi-label zero-shot learning with structured knowledge graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1576–1585.
[35] A. Gupta et al., "Generative multi-label zero-shot learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 14611–14624, Dec. 2023.
[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.
[37] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4904–4916.
[38] N. Mu, A. Kirillov, D. Wagner, and S. Xie, "SLIP: Self-supervision meets language-image pre-training," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 529–544.
[39] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.
[40] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.
[41] A. Zareian, K. D. Rosa, D. H. Hu, and S.-F. Chang, "Open-vocabulary object detection using captions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14393–14402.
[42] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–22.
[43] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[44] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–21.
[45] M. Gao et al., "Open vocabulary object detection with Pseudo bounding-box labels," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 266–282.
[46] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022.
[47] C. Ma et al., "Understanding and mitigating overfitting in prompt tuning for vision-language models," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4616–4629, Sep. 2023.
[48] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16816–16825.
[49] T.-S. Chua et al., "Nus-wide: A real-world web image database from national university of Singapore," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2009, pp. 1–9.
[50] A. Kuznetsova et al., "The open images dataset V4: Unified image classification, object detection, and visual relationship detection at scale," *Int. J. Comput. Vis.*, vol. 128, no. 7, pp. 1956–1981, 2020.
[51] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, "Deep convolutional ranking for multilabel image annotation," in *Proc. Int. Conf. Learn. Representations*, 2014, pp. 1–9.
[52] A. Veit et al., "Learning from noisy large-scale datasets with minimal supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 839–847.
[53] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–19.

Peirong Ma received the Ph.D. degree from Fudan University, Shanghai, China, in 2024. He is currently a Lecturer with the School of Computer and Electronic Information/School of Artificial Intelligence, Nanjing Normal University, Nanjing, China. His research interests include computer vision, machine learning, and pattern recognition.

Wu Ran received the doctoral degree in computer science from Fudan University, Shanghai, China, in 2024, under the supervision of Professor Hong Lu. He is currently a Postdoctoral Researcher under the supervision of Professor Chao Ma. His research interests include learning-based image restoration and novel view synthesis.

Zhiquan He received the B.S. degree in software engineering in 2022 from Fudan University, Shanghai, China, where he is currently working toward the M.S. degree with the Shanghai Key Lab of Intelligent Information Processing, School of Computer Science. His research interests include image processing, image restoration, and computer vision.

Jian Pu (Member, IEEE) received the Ph.D. degree from Fudan University, Shanghai, China, in 2014. From 2016 to 2019, he was an Associate Professor with the School of Computer Science and Software Engineering, East China Normal University, Shanghai, and a Postdoctoral Researcher with the Institute of Neuroscience, Chinese Academy of Sciences, Beijing, China, from 2014 to 2016. He is currently a Young Principal Investigator with the Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University. His current research interests include machine learning and computer vision methods for autonomous driving.



Hong Lu (Member, IEEE) received the B.Eng. and M.Eng. degrees in computer science and technology from Xidian University, Xi'an, China, in 1993 and 1998, respectively, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2005. From 1993 to 2000, she was a Lecturer and a Researcher with the School of Computer Science and Technology, Xidian University. From 2000 to 2003, she was a Research Student with the School of Electrical and Electronic Engineering, Nanyang Technological University. Since 2004, she has been with the School of Computer Science, Fudan University, Shanghai, China, where she is currently a Professor. Her current research interests include computer vision, machine learning, pattern recognition, and robotic tasks.