Hierarchical Knowledge Graph for Multilabel Classification of Remote Sensing Images

Xiangrong Zhang, *Senior Member, IEEE*, Wenhao Hong, Zhenyu Li, Xina Cheng, *Member, IEEE*, Xu Tang, *Senior Member, IEEE*, Huiyu Zhou, and Licheng Jiao, *Fellow, IEEE*

***Abstract*— Multilabel classification in remote sensing (RS) images aims to correctly predict multiple object labels in an RS image with the primary challenge of mining correlations among multiple labels. In this context, we argue that a scene can be treated as a high-level depiction of the interactions among multiple interconnected objects within the image. However, hierarchical relationships between the scene and local objects are often neglected in other state-of-the-art approaches. In this article, we consider multilabel classification as a global-to-local prediction process, whereas the scene of an image is first identified, followed by recognition of local objects in the image. To achieve this, we propose a novel hierarchical knowledge graph (HKG)-based framework for multilabel classification in RS images (ML-HKG). Specifically, we first construct a hierarchical KG to depict label correlations between scenes and objects and represent the hierarchical knowledge as interrelated scene- and object-level label embeddings. Subsequently, we generate a scene-aware enhanced feature map by recognizing scene categories in an image under the guidance of scene-level knowledge embeddings. Afterward, object-level embeddings are used to derive category-specific visual representations for final multilabel prediction. Extensive experiments on the UCM and AID datasets demonstrate the effectiveness of our framework.**

***Index Terms*— Attention mechanisms, knowledge graph (KG), multilabel classification, remote sensing (RS) images.**

I. Introduction

Remote sensing (RS) image classification has been widely applied in various geographic fields for sustainable development, including land planning [1], natural disaster prediction [2], and environment monitoring [3], [4], [5], [6]. Compared against single-label RS image scene classification methods, which consider the entire RS image as an integral unit during inference [7], [8], [9], multilabel classification methods aim to extract and recognize multiple classes of objects within a single RS image, requiring not only the ability to distinguish multiple objects but also to effectively mine the implicit correlation among them.

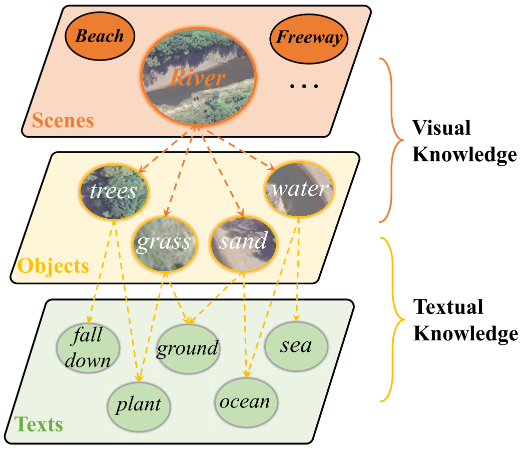


Fig. 1. Illustration of the HKG for RS images. The visual knowledge between scenes and objects can be interpreted as the compositional relationship between the global representation of an image and its local attributes, while the textual knowledge describes interrelationships among multilabel objects.

Current research on multilabel classification tasks mainly focuses on two major aspects: feature extraction of the corresponding classes from the region of interest [10] and modeling of the label correlations between multiple image labels [11]. Recently, several methods have been developed to enhance image feature representations by introducing an attention mechanism [12]. For example, cross-attention mechanisms [13], [14] are widely introduced to extract class-aware features from the region of interests by using informative label embeddings. However, these label embeddings are either randomly learned or general text concepts that fail to clearly represent the label dependencies that exist in the dataset, thus leading to limited multilabel interaction. To solve this issue, extensive works attempt to exploit the label correlation extracted from the dataset through implicit or explicit ways [15]. For instance, graph-based networks [16], [17] are introduced to explore the semantic interactions among different objects under the guidance of statistical label co-occurrence. However, due to the top–down viewpoint employed in acquiring RS imagery, different objects may exhibit visual similarities (e.g., trees and grass), thereby making it difficult to accurately differentiate based merely on co-occurrence knowledge.

In addition, given a broad field of view and complex characteristics, RS imagery often exhibits a notable

Received 3 June 2024; revised 20 September 2024; accepted 1 October 2024. Date of publication 4 November 2024; date of current version 7 November 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62276197 and Grant 62171332 and in part by the Key Research and Development Program in the Shaanxi Province of China under Grant 2019ZDLGY03-08. *(Corresponding author: Xiangrong Zhang.)*

Xiangrong Zhang, Wenhao Hong, Zhenyu Li, Xina Cheng, Xu Tang, and Licheng Jiao are with the School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: xrzhang@mail.xidian.edu.cn).

Huiyu Zhou is with the School of Computing and Mathematical Sciences, University of Leicester, LE1 7RH Leicester, U.K.

Digital Object Identifier 10.1109/TGRS.2024.3478817

hierarchical relationship from the scene to its comprising local objects [18]. As shown in Fig. 1, the global depiction of scene image river involves a combination of multiple interconnected objects, including trees, grass, sand, and water. The distinct feature of each object plays a unique role in shaping this particular scene, and these elements give rise to a scene-specific label relationship rarely observed in other RS scenes. Moreover, this intricate structure in RS images bears semblance to the interpretation in human visual perceptual learning [19], characterized by a top–down guided process that enables the recognition of complex targets through hierarchical processing, progressing from the global to the local level. While the existing approaches often take into account the interrelationship among multiple objects, the intrinsic hierarchical relationship extending from scenes to local objects within RS images is neglected. Therefore, it is necessary to develop a hierarchical method to comprehensively explore the global-to-local interactions between scenes and local objects for the precise identification of multilabel objects within RS images.

Knowledge graphs (KGs) were proposed to delineate concepts and semantic correlations through a graph structure and have witnessed rapid advancements in natural language processing [20]. Since the KG portrays human knowledge by representing entities and their semantic interrelationships as knowledge triples, it has been increasingly introduced in image classification to generate knowledge embeddings for label concepts [21] for a precise description of semantic correlations, which also exhibits great potential in representing the hierarchical knowledge within RS imagery.

In this article, we propose a hierarchical KG (HKG)-based framework for multilabel classification of remote sensing images (ML-HKG), which leverages hierarchical knowledge to represent the correlations not only within multilabel objects but also between scenes and objects. Instead of utilizing the co-occurrence knowledge to construct the label dependency matrix for exploring interactions, we transform it into the visual association between scenes and their corresponding local objects. To complement the label correlation that cannot be mined from the visual knowledge, we additionally extract object-related textual knowledge from the human common knowledge base [22] to incorporate their semantic interrelationship. By integrating visual association and textual information, an HKG is constructed to depict the label correlation and can be represented as knowledgeable label embeddings. Moreover, to enable a hierarchical semantic–visual interaction under the guidance of knowledgeable embeddings, a scene-aware feature enhancement module is introduced to inject the scene knowledge into the crude visual features, encouraging an adaptive shift in focus from global representations of the scene to the local attention of scene-associated objects, thus obtaining a scene-aware feature map. Then, a knowledge-guided feature alignment (FA) module utilizes the object-level embeddings to extract and fuse the category-specific information from the scene-aware feature for the final classification.

Overall, our contributions are as follows.

1) Inspired by hierarchical processing in human perceptual learning, we propose a global-to-local framework for multilabel image classification (MLIC), which improves the recognition of multiple objects by introducing scene-to-object relationships in RS imagery.
2) We develop an HKG integrating visual and textual knowledge, enabling a comprehensive representation of label correlations between RS scenes and multilabel objects.
3) To inject the knowledge for global-to-local learning guidance, we propose the scene-aware feature enhancement and knowledge-guided FA modules that achieve a hierarchical semantic–visual interaction.
4) Experiments on commonly used RS datasets, i.e., UCM and AID datasets, achieve a state-of-the-art result and verify the effectiveness of our model.

II. Related Work

A. MLIC Methods

MLIC of natural images that are ground level has attracted increasing attention for many years. Compared with single-label classification, the MLIC method requires better identification of multiple classes of objects appearing in one image simultaneously. To solve this issue, early work [23] leverages region proposals generated by objection detection to localize spatial features of multilabel objects for prediction. However, it neglects the label dependencies existing in MLIC [24] and, thus, attains a limited performance on label prediction.

Recently, many MLIC approaches focused on capturing the label correlation between multilabel objects. An end-to-end framework combining CNN and RNN [25] was proposed to learn a joint image-label embedding for representing semantic dependencies between labels. Chen et al. [17] introduced a semantic decoupling module to learn semantics-specific representations and a semantic interaction module to explore their interactions. Graph convolution network (GCN) [16] was first introduced to model the label correlation and to enable the semantic interaction among object embeddings by utilizing the statistical label co-occurrence. Moreover, Liu et al. [13] developed a transformer-based two-stage framework for multilabel classification, utilizing the built-in cross-attention module to adaptively explore the label correlations. Wu et al. [26] first considered the MLIC as a graph-matching problem and utilized a graph network to form structured representations for each instance and label by establishing the instance-label assignment. However, these above methods only explore the dependencies between multilabel objects, while the hierarchical correlation between scenes and objects is neglected. To tackle the issue, a scene-aware label co-occurrence module was proposed in [27] to assign an independent co-occurrence matrix for each scene category. Although this approach has obtained improved performance on ground-level images by introducing the concept of scenes, the unsupervised scene prediction mechanism within the method cannot be directly applied to remote sensing (RS) imagery considering the complex RS categories of scenes.

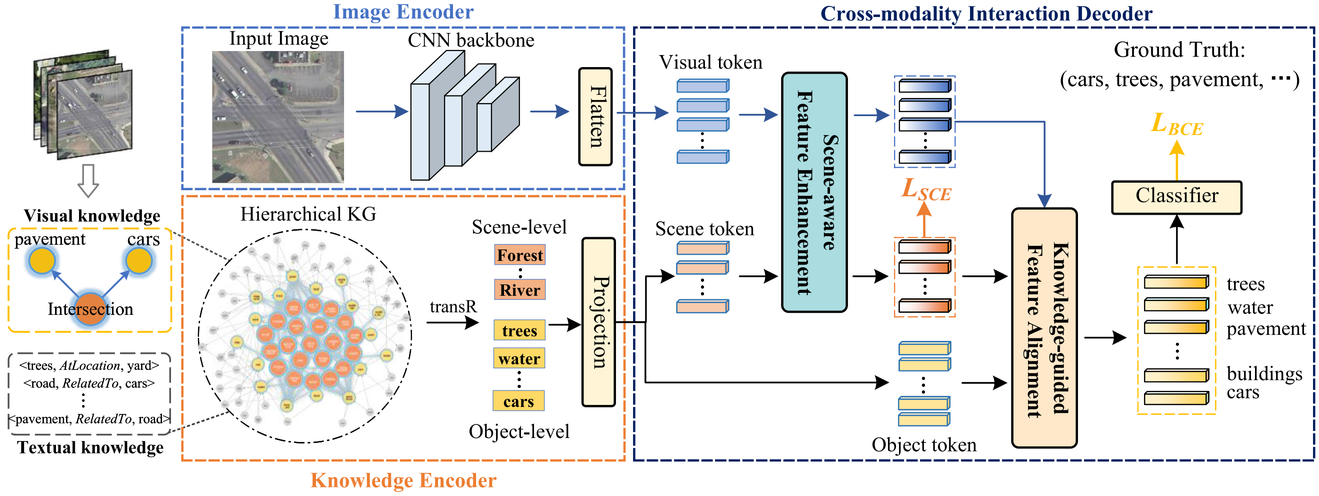


Fig. 2. Overall framework of the proposed ML-HKG. To represent the label correlation between scenes and objects, the HKG is constructed by integrating both visual knowledge from the image set and textual knowledge from the existing knowledge base. The idea of ML-HKG is to first perform scene prediction on crude global features guided by predefined scene-level knowledge embeddings to generate a scene-aware feature map, from which the object-level knowledge embeddings are utilized to extract and fuse the visual representations of corresponding objects for final multilabel classification.

B. Multilabel Classification Methods of RS Images

Compared with natural images that are ground level, RS imagery possesses a wider field of view and, thus, contains a rich variety of geological characteristics, posing a great challenge in constructing the correlations between multiple objects. To address that, Zeggada et al. [28] proposed a radial basis function neural network to yield a powerful description of RS query image by subdividing it into a grid of tiles. Hua et al. [29] introduced a classwise attention layer to generate discriminative class-specific features and a bidirectional LSTM-based subnetwork to explore label dependencies. Moreover, Tan et al. [30] first constructed a feature-based graph and a semantic graph based on low-rank representation to capture the global relationships between images, which are exploited to train a graph-based multilabel classifier. Tan et al. [31] first applied transformer to model label relationships in RS images for multilabel classification, which introduced a semantic sensitive module (SSM) to locate category-specific semantic regions and a semantic relation-building module (SRBM) to autonomously infer relationships between different categories of multilabel objects. Ma et al. [32] proposed a label-driven GCN to excavate information from the inherent label correlations, which were further enriched by utilizing a semantic enrichment module. Song et al. [33] proposed a scene-to-label prediction module to calibrate the initial multilabel prediction scores by using a scene-to-label probability matrix. While the above methods have attained competitive performance, the spatial hierarchical relationship between scenes and local objects within RS images still remains to be explored.

C. KGs for Representing Label Correlations

KG was officially introduced by Google [34] in 2012, to store structured facts and concepts in the physical world in a symbolic form and to visualize them as a multirelational graph [35]. As defined in the previous work [20], a KG is an aggregation of entities and their associated relations, regarded as nodes and different types of edges, respectively. Since the KG comprised of triples provides rich external human knowledge, which is hard for machines to learn from images, it is now increasingly applied in computer vision tasks [21], [36]. Lin et al. [37] were the first to introduce a concept graph into the multilabel classification of RS images to describe label correlations in RS images. After obtaining label-related entities from the knowledge base ConceptNet [22] and using them to reconstruct a subgraph, a concept attention graph neural network was then proposed to extract label correlation features from the concept graph, which were sent to a binary classifier with the image features. However, the concept graph only utilized the label-associated entities, discarding the wide variety of relationships between different entities that represent valuable prior knowledge within the knowledge triples, and the label correlation features were fixed, since they were obtained by only integrating information from predefined features of label nodes. In addition, while the method depicted human knowledge, it neglected much visual knowledge like label co-occurrence in specific aerial datasets, which is certainly essential during multilabel inference. In fact, in different RS scene settings, the label correlations between objects are considered to bear some slight difference. For instance, buildings and trees are more closely related to the scene of medium residential (since they are both important elements constituting this type of scene images), while less relevant in the scene of forest.

Moreover, KGs have been successfully and widely utilized in knowledge representation learning (KRL), which embeds interrelated entities into low-dimensional vectors while also maintaining their semantic relationships [38], [39]. Subsequently, KRL has become one of the most popular research areas, and researchers have proposed many models to embed entities and relations in KGs [40] into pretrained knowledgeable vectors, which could be leveraged for further applications. Li et al. [41] first applied KRL to the RS field by constructing an RS KG based on the domain prior knowledge from human experts, which improves the semantic representation ability of RS-oriented scene categories in the zero-shot RS image scene classification problem.

To the best of our knowledge, our work is the first to incorporate both visual and textual knowledge to build a KG and to explore the hierarchical label correlations among scenes and multilabel objects within RS imagery utilizing KRL methods.

III. Methodology

Given an RS image set $I = \{x_1, x_2, \dots, x_n\}$ containing n images, a scene label set $S = \{s_1, s_2, \dots, s_K\}$ containing K categories of scenes, and an object label set $L = \{l_1, l_2, \dots, l_M\}$ containing M categories of objects, single-label scene classification aims to identify the scene category to which the ith image $x_i \in I$ belongs, while multilabel classification aims to predict whether each object is present in the image of sth scene x_i^s, with the object labels of x_i^s represented as $y_i = \{y_{i1}, \dots, y_{iM}\}$, where $y_{im} = 1$ if the mth object label is present, $m = 0, 1, \dots, M$.

Fig. 2 illustrates the framework of the proposed ML-HKG, which consists of three main components: a knowledge encoder that encodes the constructed HKG into scene and object knowledge embeddings, an image encoder that processes the input image to obtain a crude visual feature, and a cross-modality interaction decoder module that enables the scene-aware feature enhancement and visual–semantic alignment guided by knowledge embeddings. The idea of ML-HKG is to first perform scene prediction on the crude visual features by injecting scene-level knowledge embeddings to generate a scene-aware feature map, from which the object embeddings are utilized to extract and fuse the visual representations of corresponding objects for the final classification.

A. HKG Construction and Representation

KGs have been introduced to store structural facts and concepts that depict the knowledge of the real world. As defined in [20], a KG is an aggregation of multiple triples $\mathrm{T} = \{\langle h, r, t\rangle | h, t \in \varepsilon, r \in R\}$ given the entity set ε and the relationship set R, where h and t denote the head entity and tail entity, respectively, and r denotes their specific interrelationship. Based on the above definition, we construct the visual occurrence knowledge between scenes and objects to represent the label dependencies within the dataset. Specifically, given a multilabel RS image dataset with each image also labeled with a specific scene, we treat the images of scene s_i as an individual set to record all categories of multilabel objects that have appeared. Then, the scenes and their corresponding objects are connected in the form of knowledge triples $\langle s_k, \text{AssociatedWith}, l_m\rangle$, where the scene s_k and object l_m are defined as entities and their interrelationship is unified as AssociatedWith to represent the visual co-occurrence between RS scenes and objects. For instance, for a specific scene river in the UCM dataset, we record all the objects (e.g., trees, water, and sand) that have appeared in the images of river and construct the triples, such as $\langle$river, AssociatedWith, water$\rangle$.

Moreover, to complement the label correlation that cannot be mined from the visual knowledge, we additionally extract the object-related textual knowledge from the knowledge base ConceptNet [22] without introducing additional information unrelated to RS scenarios. In detail, we follow [37] to extract relevant knowledge triples based on both the explicit and implicit correlations between two objects, in which the explicit knowledge triple set $T_{\text{exp}} = \{\langle l_p, r, l_q\rangle | l_p, l_1 \in L\}$ indicates that l_p and l_q are directly related to the common world, and the implicit triple set $\mathrm{T}_{\text{imp}} = \{\langle l_p, r, e\rangle, \langle e, r, l_q\rangle | l_p, l_1 \in L\}$ indicates l_p and l_q are indirectly related through a shared entity. For example, by searching for relationships between object label trees and grass from the knowledge base ConceptNet, we obtain the explicit knowledge triple $\langle$grass, AtLocation, trees$\rangle$ and the implicit triples $\langle$grass, RelatedTo, plant$\rangle$ and $\langle$plant, RelatedTo, trees$\rangle$. Note that [37] only utilized the definition to extract object-related entities and discarded the predefined relationships between different entities that contain valuable prior knowledge in the form of knowledge triples, while our ML-HKG retains these triples with different types of relationships to distinguish between different object-level labels and cluster them and the above visual knowledge triples. By integrating both visual and textual knowledge triples, an HKG is, therefore, constructed to represent correlations between scenes and multilabel objects.

Then, to embed the structural HKG into the semantic space for subsequent cross-modality interaction, we adopt transR [42] to transform entities and relationships into feature vectors. Specifically, for each knowledge triple $\langle h, r, t\rangle$, the embeddings of the head entity and the tail entity are initialized as $\boldsymbol{e}_h, \boldsymbol{e}_t \in \mathbb{R}^{d_e}$, and the relation embedding is set as $\boldsymbol{r} \in \mathbb{R}^{d_r}$, where d_e and d_r denote the dimension of the entities and the relationship, respectively. For each relation r, we utilize a projection matrix $\boldsymbol{M}_r \in \mathbb{R}^{d_e \times d_r}$ to project both the head and tail entities from entity space into relation space

$$\boldsymbol{h}_r = \boldsymbol{e}_h \boldsymbol{M}_r \tag{1}$$

$$\boldsymbol{t}_r = \boldsymbol{e}_t \boldsymbol{M}_r. \tag{2}$$

With the relation-specific mapping matrix, the head entity and tail entity in a triple can be embedded into semantic vectors $\boldsymbol{h}_r, \boldsymbol{t}_r \in \mathbb{R}^{d_r}$, and the relationship between them can be satisfied by the following objective function:

$$f_r(h, t) = \|\boldsymbol{h}_r + \boldsymbol{r} - \boldsymbol{t}_r\|_2^2. \tag{3}$$

By optimizing the objective function in (3), we can embed both the visual and textual knowledge into semantic representations, thus obtaining the knowledgeable scene embeddings $\boldsymbol{\varepsilon}^s \in \mathbb{R}^{K \times d_e}$ and object embeddings $\boldsymbol{\varepsilon}^o \in \mathbb{R}^{M \times d_e}$, where K and M correspond to the number of scene categories and object categories, respectively.

After obtaining the pretrained knowledge embeddings, a shared linear projection is applied to map them into visual space for the subsequent cross-modality interaction

$$\boldsymbol{S} = \boldsymbol{\varepsilon}^s \boldsymbol{w}_1 \tag{4}$$

$$\boldsymbol{E} = \boldsymbol{\varepsilon}^o \boldsymbol{w}_1 \tag{5}$$

where $\boldsymbol{w}_1 \in \mathbb{R}^{d_e \times C}$ is a learnable weight matrix for the linear layer and C is the dimension of visual space.

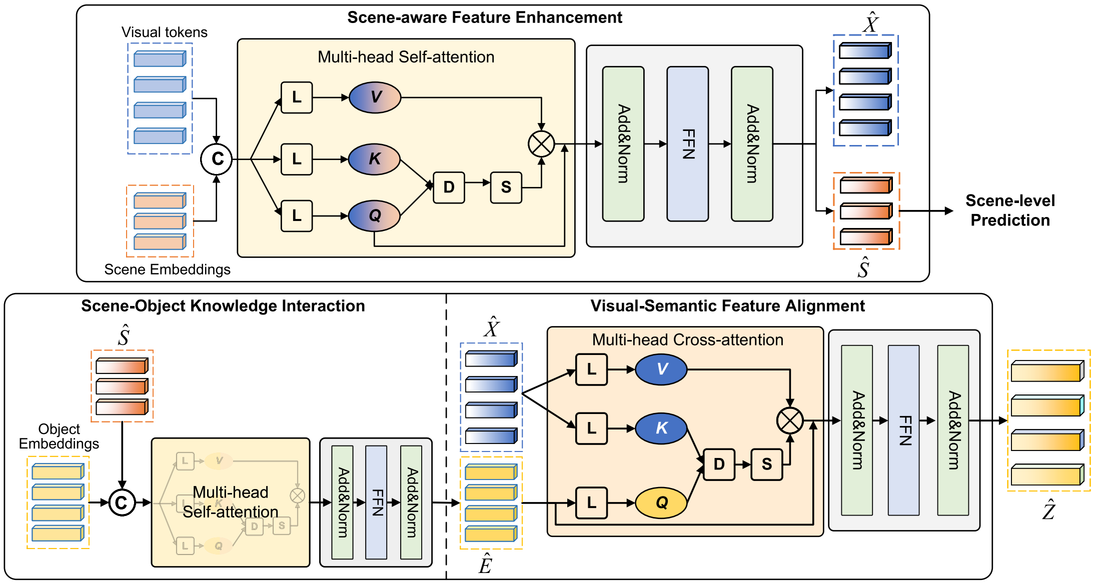

Scene-aware Feature Enhancement
Visual tokens
Multi-head Self-attention
L
V
K
Q
C
D
S
Add&Norm
FFN
Add&Norm
$\hat{X}$
$\hat{S}$
Scene-level Prediction
Scene Embeddings
Scene-Object Knowledge Interaction
Object Embeddings
Multi-head Self-attention
Visual-Semantic Feature Alignment
Multi-head Cross-attention
$\hat{E}$
$\hat{Z}$

Fig. 3. Detail of the proposed cross-modality interaction decoder. The scene-aware feature enhancement module takes visual tokens and scene-level knowledge embeddings as the input and outputs the scene-aware visual features and updated scene embeddings. The knowledge-guided FA module consists of the scene-object KI component for imagewise knowledge updating and the visual–semantic FA component for cross-modality feature fusion.

B. Scene-Aware Feature Enhancement

To acquire visual representations of multiple objects, previous methods [16] tend to directly exploit the crude visual features extracted from a pretrained CNN network to obtain the discriminative object information. However, in the crude feature map, the visual features at different regions are obtained by the convolution kernels' independent processing, which lacks the global perception of the scene image to reflect the explicit interaction of multilabel objects under different positions. Also, it is noted that in the human visual perception learning system, there is a top–down processing procedure that progresses from the coarse global information perception to more fine-grained local discrimination, and this hierarchical process can improve the human cognition ability of surrounding objects. Therefore, it is necessary to perform preliminary scene-level recognition of crude visual features to obtain discriminative visual features of multilabel objects in the scene-aware context.

Recently, prompt engineering has been widely applied in computer visual tasks [43] to fine-tune the pretrained model to generate task-specific outputs for downstream fields by devising trainable or nontrainable prompt tokens and inserting them into the frozen model. Chen et al. [44] utilized textual features of class labels extracted from the text encoder as the semantic prompts, which were injected into the visual feature extractor to adaptively tune the feature extraction to class-specific features.

Inspired by the idea of prompting [45], we take the scene-level embeddings as the global knowledge prompts and inject them into the crude visual features through our scene-aware feature enhancement (SEF) module (Fig. 3), thus encouraging the shift in feature extraction from global representation to local attention on objects. Note that our knowledge prompts are fixed embeddings pretrained from our knowledge encoder, and other components in our ML-HKG are involved in training, which is different from the former methods.

Given the input image x, after obtaining the visual tokens $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{HW}\}$ by flattening the crude spatial feature maps, we concatenate them with the knowledgeable scene embeddings, which are fed together into a standard multihead self-attention (MHSA) following [46]:

$$[\boldsymbol{Z}_X, \boldsymbol{Z}_S] = \mathrm{MHSA}([\boldsymbol{X}, \boldsymbol{S}]). \tag{6}$$

Specifically, for the mth region visual token $\boldsymbol{x}_i$, its attention weight of knowledge absorption from the nth scene embedding $\boldsymbol{s}_i$ is formulated as follows:

$$\alpha_{mn} = \frac{\exp\left(\left(\boldsymbol{x}_m \boldsymbol{W}_q^e\right)\left(\boldsymbol{s}_n \boldsymbol{W}_k^e\right)^T \Big/ \sqrt{d}\right)}{\sum \exp\left(\left(\boldsymbol{X} \boldsymbol{W}_q^e\right)\left([\boldsymbol{X}, \boldsymbol{S}]\boldsymbol{W}_k^e\right)^T \Big/ \sqrt{d}\right)} \tag{7}$$

where $\boldsymbol{W}_q^e$ and $\boldsymbol{W}_k^e$ are learnable weight metrics and d is a scaling factor. As a result, the overall scene-level knowledge that $\boldsymbol{x}_i$ absorbs is the weighted sum of all scene embeddings

$$\boldsymbol{Z}_{x_m \leftarrow S} = \sum_{n=1}^{K} \alpha_{mn} \left(\boldsymbol{s}_n \boldsymbol{W}_v^e\right) \tag{8}$$

where $\boldsymbol{W}_v^e$ is a learnable weight matrix. Through our SFE module, the crude visual feature can adaptively select and absorb the most related scene knowledge among a set of scene candidates and focus on the local features of dominant scene-specific objects, and the scene knowledge embeddings are enabled to extract the relevant discriminative representation from the global visual features.

Then, a shared feedforward network (FFN) is applied for intermodality information integration, thus obtaining the scene-aware visual features and the updated scene representations for scene category prediction and further FA

$$\hat{\boldsymbol{X}} \leftarrow \mathrm{FFN}(\boldsymbol{X} + \boldsymbol{Z}_X) + \boldsymbol{X} \tag{9}$$

$$\hat{\boldsymbol{S}} \leftarrow \mathrm{FFN}(\boldsymbol{S} + \boldsymbol{Z}_S) + \boldsymbol{S} \tag{10}$$

where FFN denotes the feedforward network layer, and $\boldsymbol{Z}_X$ and $\boldsymbol{Z}_S$ are the outputs from the cross-modality attention module.

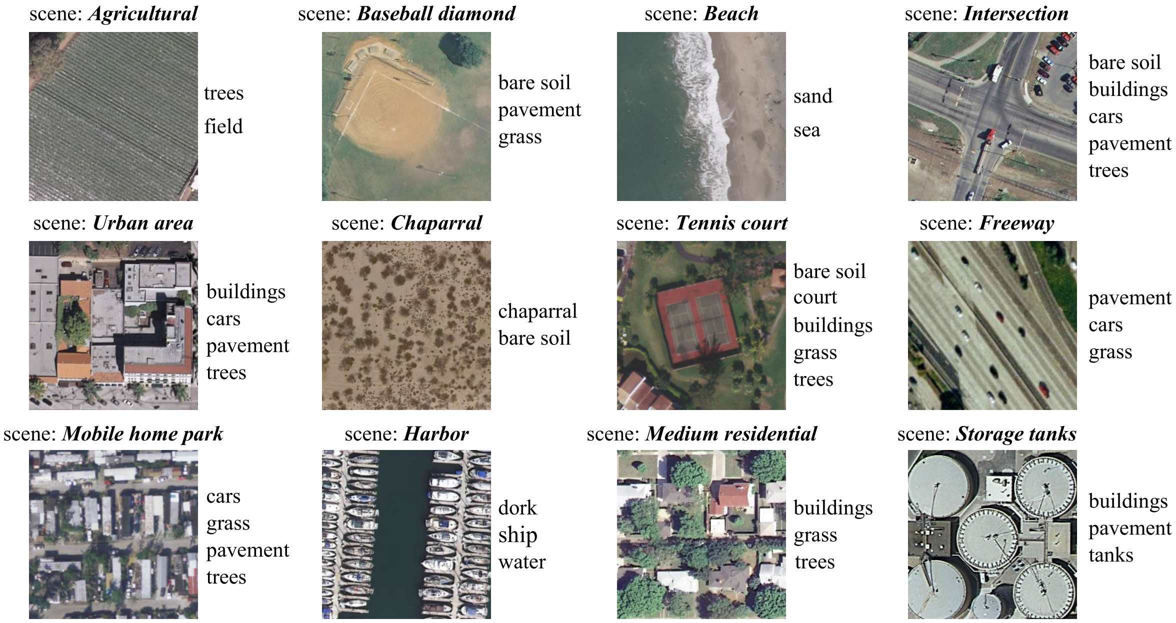


Fig. 4. Example images of the UCM dataset, with the corresponding scene categories and multilabel objects.

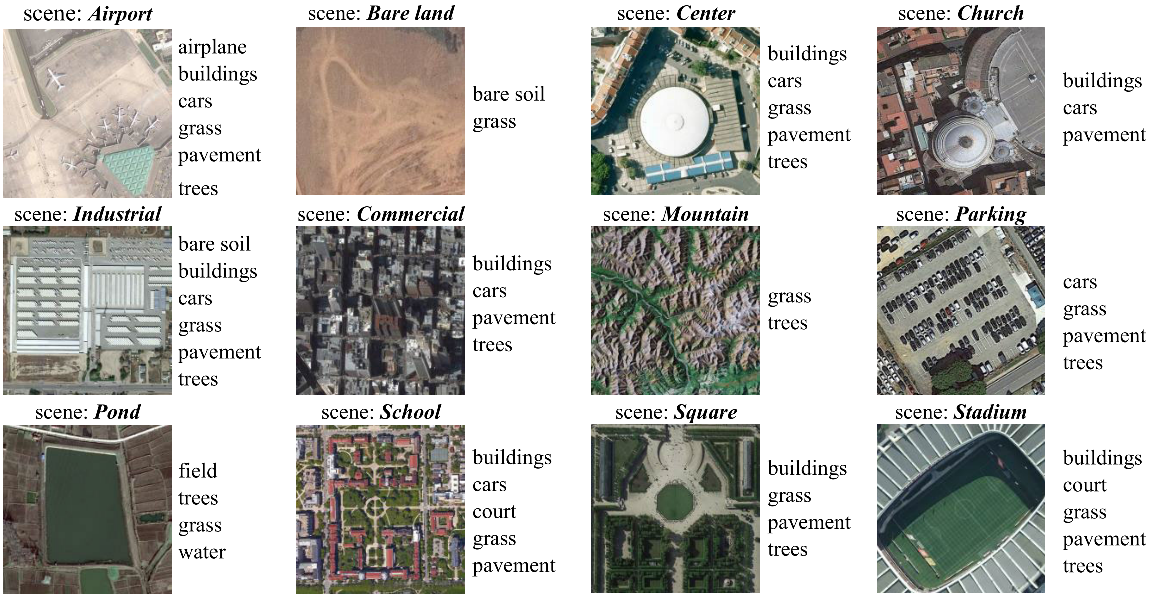


Fig. 5. Example images of the AID dataset, with the corresponding scene categories and multilabel objects.

C. Knowledge-Guided FA

1) Scene-Object KI: Since the updated global representation integrated by the scene-level embeddings contains discriminative information about the scene of the input image, it also reveals the presence and interaction of the dominant local objects. Hence, to effectively transfer the updated samplewise global knowledge onto multilabel objects and selectively focus attention on exploring the existence of specific objects after perceiving the scene label, we combine the scene embeddings output from the SFE module with the pretrained object embeddings to further propagate information on the hierarchical correlations

$$[\boldsymbol{Z}_E, \boldsymbol{Z}_S] = \text{MHSA}([\boldsymbol{E}, \hat{\boldsymbol{S}}]). \tag{11}$$

In addition, an FFN with two linear layers and a nonlinear activation in between is applied to output the updated object knowledge embeddings $\hat{\boldsymbol{E}} \leftarrow \text{FFN}(\boldsymbol{E} + \boldsymbol{Z}_E) + \boldsymbol{E}$.

2) Visual–Semantic FA: After obtaining the scene-aware visual feature and the updated object-level embeddings, we resort to the cross-attention mechanism [13] to enable the cross-modality feature interaction and alignment. In detail, we take the object-level knowledge embeddings as queries to incorporate visual information of the corresponding objects from the enhanced visual features

$$\boldsymbol{Z} = \text{softmax}\left(\frac{\hat{\boldsymbol{E}}\boldsymbol{W}_q^d \left(\hat{\boldsymbol{X}}\boldsymbol{W}_k^d\right)^T}{\sqrt{d}}\right)\hat{\boldsymbol{X}}\boldsymbol{W}_v^d \tag{12}$$

$$\hat{\boldsymbol{Z}} \leftarrow \text{FFN}(\hat{\boldsymbol{E}} + \boldsymbol{Z}) + \hat{\boldsymbol{E}} \tag{13}$$

TABLE I

NUMBER OF DIFFERENT OBJECT-LEVEL LABELS IN THE UCM DATASET

Object label	Number	Object label	Number
airplane	100	mobile home	102
bare soil	718	pavement	1300
buildings	691	sand	294
cars	886	sea	100
chaparral	115	ship	102
court	105	tanks	100
dock	100	trees	1009
field	103	water	203
grass	975		

TABLE II

NUMBER OF DIFFERENT OBJECT-LEVEL LABELS IN THE AID DATASET

Object label	Number	Object label	Number
airplane	99	mobile home	2
bare soil	1475	pavement	2328
buildings	2161	sand	259
cars	2026	sea	221
chaparral	112	ship	284
court	344	tanks	108
dock	271	trees	2406
field	214	water	852
grass	2295		

where $\boldsymbol{W}_q^d$, $\boldsymbol{W}_k^d$, and $\boldsymbol{W}_v^d$ are learnable weight matrices, and $\hat{\boldsymbol{Z}} \in \mathbb{R}^{M \times C}$ denotes the updated multilabel representations.

Thus, our KFA module can effectively localize the regions most relevant to each object from the enhanced features under the guidance of object-level knowledge and generate the label representations integrating both semantic knowledge and visual information for final multilabel classification.

D. Learning Objective

1) Scene-Aware Cross-Entropy Loss: Given an input image x, to encourage the SFE module to accurately perceive the scene of the image and focus attention on scene-specific objects, we utilize the updated scene representations in Section III-B to predict the scene category under the scene-level supervision. In detail, the probabilities of scenes for image x are obtained by computing the similarity scores between the scene embeddings $\hat{\boldsymbol{S}}$ and the corresponding scene prototypes followed by a softmax function. Then, the cross-entropy loss is adopted to calculate the scene-aware loss

$$L_{\text{sce}} = -\log \frac{\exp\left(\hat{\boldsymbol{s}}_k^{\mathrm{T}} \boldsymbol{w}_k\right)}{\sum_{k=1}^{K} \hat{\boldsymbol{s}}_k^{\mathrm{T}} \boldsymbol{w}_k} \tag{14}$$

where $\boldsymbol{s}_k$ denotes the kth scene embedding and $\boldsymbol{w}_k \in \mathbb{R}^C$ is the learnable prototype of the kth scene class.

2) Multilabel Classification Loss: For the multilabel classification, we feed the aligned object embeddings $\hat{Z}$ into a set of independent binary classifiers to predict the labels separately with a sigmoid operation and adopt the cross-entropy loss to constrain the training process

$$p_i = \text{Sigmoid}\left(\hat{\boldsymbol{z}}_i^{\mathrm{T}} \boldsymbol{W}_i + \boldsymbol{b}_i\right) \tag{15}$$

where $\hat{\boldsymbol{z}}_i$ is the ith object embedding, and $\boldsymbol{W}_i \in \mathbb{R}^C$ and $\boldsymbol{b}_i \in \mathbb{R}$ are the weight and bias for the ith object, respectively,

$$L_{\text{bce}} = \frac{1}{M} \sum_{i=1}^{M} y^i \log(p_i) + (1 - y^i) \log(1 - p_i). \tag{16}$$

Combining the scene-level and the object-level supervision loss, the final sample-level loss is formulated as follows:

$$L = L_{\text{bce}} + \lambda L_{\text{sce}} \tag{17}$$

where λ is the balance parameter between the two losses.

IV. EXPERIMENT

A. Datasets

1) UCM Dataset: The UCM dataset [48] serves as a common benchmark dataset and is widely used for evaluating both single-label scene classification and multilabel object classification of RS imagery. The dataset contains 2100 aerial images divided into 21 categories of scenes, including agricultural, airport, baseball diamond, beach, urbanarea, chaparral, dense residential, forest, freeway, golfcourse, harbor, intersection, medium residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks, and tennis course, each of which contains 100 images of size 256 × 256 × 3. The images are further assigned 17 categories of multilabel objects [49], which are airplane, bare soil, buildings, cars, chaparral, court, dock, field, grass, mobile home, pavement, sand, sea, ship, tanks, trees, and water, with 3.3 object labels per image on average, and the total number of each category of object is shown in Table I. Fig. 4 shows some examples of the images with different scenes and the corresponding multilabel objects. Following [37], we randomly select 80% of images evenly from each scene category as the training set and the rest 20% as the testing set.

2) AID Dataset: The AID dataset [50] is another popular benchmark for multilabel classification of RS images, which contains 3000 images classified into 30 scene categories, including airport, bare land, baseball field, beach, bridge, center, church, commercial, dense residential, desert farmland, forest, industrial, meadow, medium residential, mountain, park, parking, playground, pond, port, railway station, resort, river, school, sparse residential, square, stadium, storagetanks, and viaduct, and the total number of each category of object is shown in Table II. The scene images can be subdivided into 17 multilabel object categories, which are consistent with the labels of the UCM dataset mentioned above, with each image having 5.2 object labels on average. Fig. 5 shows some examples of the images with different scenes and the corresponding multilabel objects. During the whole process, 80% of the images are randomly selected as the training set and 20% for the network testing.

B. Evaluation Metrics

For a fair comparison with other state-of-the-art approaches, we report the average of example-based metrics and the label-based metrics as the evaluation metrics. The example-based

TABLE III

COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE UCM DATASET (%)

Method	EF_1	EF_2	EP	ER	LF_1	LF_2	LP	LR
ResNet-50 [47]	79.68	80.58	80.86	81.95	83.59	80.76	88.78	78.98
ResNet-RBFNN [28]	80.58	82.47	79.92	84.59	84.95	84.21	86.21	83.72
CA-ResNet-BiLSTM [29]	81.47	85.27	77.94	89.02	85.18	84.63	86.12	84.26
ML-GCN [16]	87.01	86.08	90.95	86.15	88.67	86.73	92.80	85.61
SSGRL [17]	88.37	88.70	89.64	89.36	90.31	89.53	91.83	89.05
ResNet50-SR-Net [31]	88.67	89.11	87.96	89.40	-	-	93.52	91.51
Q2L [13]	88.83	89.99	88.60	91.29	90.88	91.07	91.06	91.33
ML-CG [37]	87.35	87.49	89.35	88.14	85.71	83.94	90.53	82.73
SALGL [27]	87.67	87.33	90.15	87.52	89.07	88.53	90.52	88.33
ML-HKG	**90.49**	**90.94**	**91.28**	**91.62**	**92.88**	**92.13**	**94.38**	**91.69**

precision (EP) and recall (ER) scores are computed as follows:

$$\mathrm{EP} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i} \tag{18}$$

$$\mathrm{ER} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i} \tag{19}$$

where TP_i and FP_i denote the numbers of true positive labels and false positive labels of the ith image, FN_i is the number of false negative labels predicted on the ith image, and N is the overall number of predicted images. Then, we adopt exampled-based F scores (EF_1 and EF_2) to evaluate the overall performance of the model on each sample based on balancing EP and ER

$$\mathrm{EF}_\beta = \frac{1}{N}\sum_{i=1}^{N}\left(1+\beta^2\right)\frac{\mathrm{EP}_i \times \mathrm{ER}_i}{\beta^2 \times \mathrm{EP}_i + \mathrm{ER}_i}, \quad \beta = 1, 2 \tag{20}$$

where EP_i and ER_i represent the ith example-based precision and recall scores.

In addition, the label-based precision (LP) and recall (LR) scores are calculated as follows:

$$\mathrm{LP} = \frac{1}{C}\sum_{l=1}^{C}\frac{\mathrm{TP}_l}{\mathrm{TP}_l + \mathrm{FP}_l} \tag{21}$$

$$\mathrm{LR} = \frac{1}{C}\sum_{l=1}^{C}\frac{\mathrm{TP}_l}{\mathrm{TP}_l + \mathrm{FN}_l} \tag{22}$$

where TP_l is the number of correctly predicted positive images, FP_l denotes the number of falsely predicted positive images, FN_l represents the number of falsely predicted negative images on the lth object label, and C denotes the total number of object-level labels. In addition, we also calculate the average of label-based F scores (LF_1 and LF_2) to attain a more balanced evaluation standard considering both LP and LR

$$\mathrm{LF}_\beta = \frac{1}{C}\sum_{l=1}^{C}\left(1+\beta^2\right)\frac{\mathrm{LP}_l \times LR_l}{\beta^2 \times \mathrm{LP}_l + LR_l}, \quad \beta = 1, 2 \tag{23}$$

where LP_l and LR_l represent label-based precision and recall on the lth object label.

While example-based metrics are more concerned with the performance of the model at the sample level, label-based metrics reflect the performance of the model on each specific label. These metrics provide a comprehensive view of the model's performance from both example-based and label-based perspectives. In addition, we also compute and report the mean average precision (mAP) for performance evaluation in the ablation study.

C. Implementation Details

We adopt ResNet-50 [47] pretrained on ImageNet [51] as the CNN backbone and resize the input images to 448 × 448 in both the training and testing phases. Following [37], we select 80% of images evenly from each scene category as the training set and the rest as the testing set. The dimension of the KG embeddings is fixed to 50, and the weight λ is set as 1.0 throughout all experiments. During training, we adopt RandAugment [52] and Cutout [53] strategies for data augmentation, and the network is trained for 50 epochs in total, with an initial learning rate of 10^{-5} and AdamW [54] as the optimizer. We implemented the network on PyTorch with a single NVIDIA GeForce RTX 3080 GPU.

D. Comparisons With the State of the Arts

In this section, we evaluate and present the performance results of our proposed ML-HKG conducted on the above datasets. First, we compare the results of our ML-HKG with the baseline model ResNet-50 to demonstrate the effectiveness of our framework. Then, the proposed method is compared with other state-of-the-art approaches. In addition, some graph-based and scene-aware MLC methods are also incorporated for a fair comparison. Furthermore, ablation studies are performed to verify the effectiveness of the modules in our framework and discuss the effects of hyperparameters.

1) Comparison on UCM: We compare our method with other state-of-the-art methods on the UCM benchmark, including ResNet-50 [47], ResNet-RBFNN [28], CA-ResNet-BiLSTM [29], ML-GCN [16], SSGRL [17], ResNet50-SR-Net [31], Q2L [13], ML-CG [37], and SALGL [27], and the experimental results are shown in Table III. For a fair comparison, the resolution of input images is resized to 448 × 448, and the optimal scores are highlighted in

TABLE IV
COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE AID DATASET (%)

Method	EF_1	EF_2	EP	ER	LF_1	LF_2	LP	LR
ResNet-50 [47]	87.43	85.98	89.95	85.04	61.36	56.97	79.90	54.87
ResNet-RBFNN [28]	83.77	85.87	82.84	88.32	62.30	68.30	60.85	70.45
CA-ResNet-BiLSTM [29]	87.63	88.03	89.03	88.99	71.88	67.98	79.50	65.60
ML-GCN [16]	89.60	89.67	91.39	90.31	75.99	72.74	82.10	70.72
SSGRL [17]	89.14	89.12	91.23	89.69	74.46	72.28	78.41	70.89
ResNet50-SR-Net [31]	89.97	90.30	89.42	90.52	-	-	**87.24**	**82.25**
Q2L [13]	89.82	90.76	90.19	92.03	75.68	74.46	80.21	73.96
ML-CG [37]	88.04	87.86	90.42	88.27	74.88	72.44	81.82	71.09
LD-GCN [32]	90.93	-	**92.81**	89.06	77.49	-	85.14	71.22
SALGL [27]	88.89	89.60	89.74	90.63	76.04	75.26	82.13	75.11
ML-HKG	**91.34**	**91.63**	92.54	**92.28**	**79.31**	**78.56**	82.82	78.25

bold. It can be observed that our model achieves the best performance in terms of all metrics. For the sample-based metrics, our approach achieves 91.28% in EP and 91.62% in ER, improving by 1.93% and 3.48% compared with ML-CG, the concept graph-based method for MLIC. Although the latter also explored the label correlations by constructing a concept graph, it discarded the prior structural knowledge within the original knowledge triples, and the label correlation features were obtained by only integrating information from predefined label nodes without the visual–semantic interaction. By incorporating both visual and textual knowledge triples in our global-to-local framework to model the hierarchical relationships neglected by ML-CG, our method can better capture the label relevance of multiple objects and explore the imagewise semantic–visual feature interaction, thus making better prediction results.

As for the more balanced evaluation metrics, the EF_1 score and EF_2 of our ML-HKG are 90.49% and 90.94%, respectively, 2.82% and 3.61% higher than SALGL. While the latter also introduced the idea of scenes by maintaining a co-occurrence matrix for each scene, due to its unsupervised scene detection module, which perceives the category without actual scene-level labels, it is less effective when applied to RS images, which contain rich geographic features and a wide variety of scenes. In contrast, by introducing scene-level supervision through our SFE module, our approach can identify the real scene of RS images and characterize the correlation between scenes and objects with more accuracy, leading to better multilabel classification performance.

Meanwhile, in terms of the label-based metrics, our method also attains the optimal results with 94.38 % in LP and 91.69% in LR, demonstrating that our model can maintain both high precision score and recall score. Compared with ML-GCN, which introduced GCN to explore object-level label correlations, our proposed method shows an improvement of 1.58% in LP and 6.08% in LR, validating the superiority of the HKG in representing label correlations. In conclusion, the experimental results show that our ML-HKG can achieve optimal performance by exploring the hierarchical label correlations among scene and object labels through both the proposed HKG and the hierarchical visual–semantic interaction process.

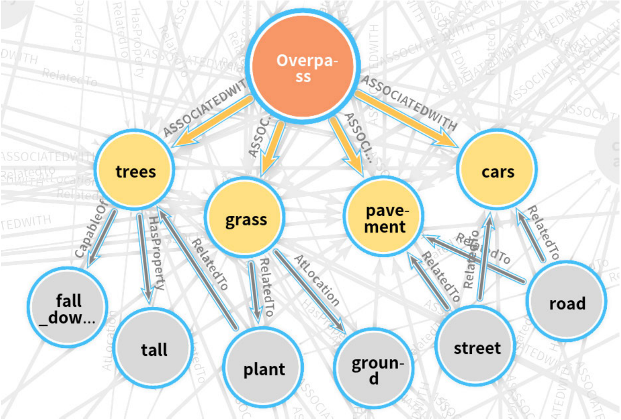


Fig. 6. Part of the visualization results of our HKG on the UCM dataset (HKG-UCM).

2) Comparison on AID: We compare our proposed ML-HKG with the state-of-the-art methods, including ResNet-50 [47], ResNet-RBFNN [28], CA-ResNet-BiLSTM [29], ML-GCN [16], LD-GCN [32], and SSGRL [17]. Table IV shows the qualitative results of our method on the AID dataset and the comparison with other state-of-the-art approaches, from which we can see that our method achieves the highest scores on most of the metrics, especially the example-based metrics. As for the example-based metrics, while our ML-HKG achieves the suboptimal result of 92.54% on EP, 0.27% lower compared with that of LD-GCN, it still improves by 3.22% on ER, and other metric values are also higher than the latter approach. This indicates that although the inherent label correlations can be semantically enriched using a label-driven graph network, our ML-HKG can achieve higher performance through a hierarchical visual–semantic interaction. Moreover, SALGL only attains 76.04% in LF_1 and 75.26% in LF_2, 3.27% and 3.3% lower than our method, which further verifies the complexity of scene features in RS images and the importance of scene-level supervision during scene perception.

Furthermore, in terms of label-based evaluation metrics, our ML-HKG achieves excellent results, with 79.31% of LF_1 score and 78.56% of LF_2 score, which is 3.63% and 4.1% higher than Q2L, which only utilized visual features during the whole training process and ignored the importance of textual knowledge. Note that ResNet-50-SRNet [31] attains the

TABLE V

EFFECT OF MODULES ON THE PERFORMANCE OF OUR ML-HKG FRAMEWORK (%)

Module			UCM			AID		
SFE	KI	FA	EF_1	LF_1	mAP	EF_1	LF_1	mAP
			87.38	89.37	96.00	88.71	76.02	83.70
		✓	88.67	90.67	96.18	89.73	76.48	84.68
	✓	✓	89.35	91.49	96.81	89.73	77.51	84.82
✓		✓	89.65	91.71	96.83	90.07	79.10	84.92
✓	✓	✓	**90.49**	**92.88**	**97.28**	**91.34**	**79.31**	**85.11**

optimal scores in LP and LR, because it successfully predicted the object label mobile home that shows up only once both in the training set and the testing set. In contrast, our approach focuses more equally on the presence of multilabel objects in all examples of the dataset, achieving 92.54% in EP and 92.28% in ER, which are 3.12% and 1.76% higher than the aforementioned method, respectively.

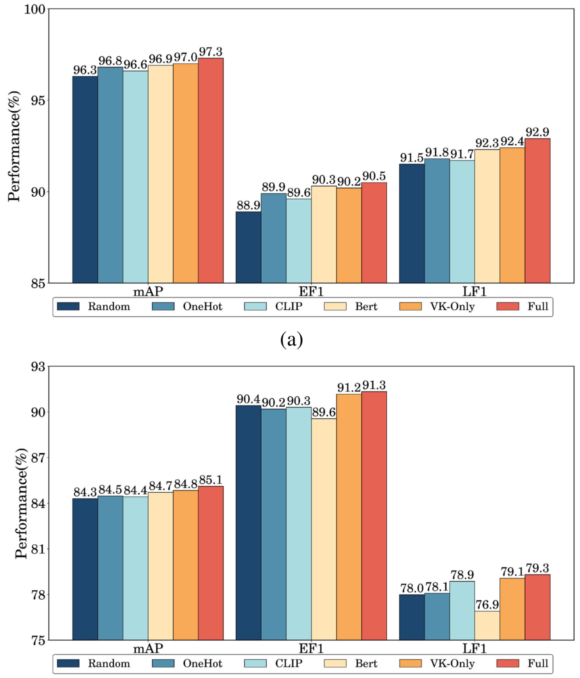
E. HKG Visualization

In this section, we provide details about our HKG and visualize the hierarchical label correlations among RS scenes. By collecting the visual co-occurrence to form the knowledge triples between RS scenes and multilabel objects and extracting the textual knowledge triples that contain interrelated objects, our ML-HKG can construct an HKG that comprehensively represents the label correlations in RS datasets. For the UCM dataset, our proposed HKG contains 218 knowledge triples, consisting of 80 categories of entities (i.e., *scenes*, *objects*, and object-related neighbors) and 10 types of relationships (e.g., *IsA*, *RelatedTo*, and *AssociatedWith*). For the AID dataset, it has 372 triples, including 89 entities and 10 relationships.

Fig. 6 shows the visualization of our HKG-UCM subgraph, which depicts the hierarchical relationships between the scene overpass and interrelated multilabel objects, where the orange node indicates the scene-level label, the yellow nodes represent object-level labels, and the gray nodes indicate other entities. In Fig. 6, the scene node overpass is directly connected to multilabel objects, including trees, grass, pavement, and cars, with the specific edge AssociatedWith to represent their relevance in the visual knowledge. Moreover, the object-level nodes are implicitly correlated through shared neighbor nodes in the textual knowledge. For example, the objects pavement and cars are both connected to the nodes street and road with the same relationship RelatedTo, and trees shares the same attribute tall with other unlisted RS objects.

F. Performance Analysis

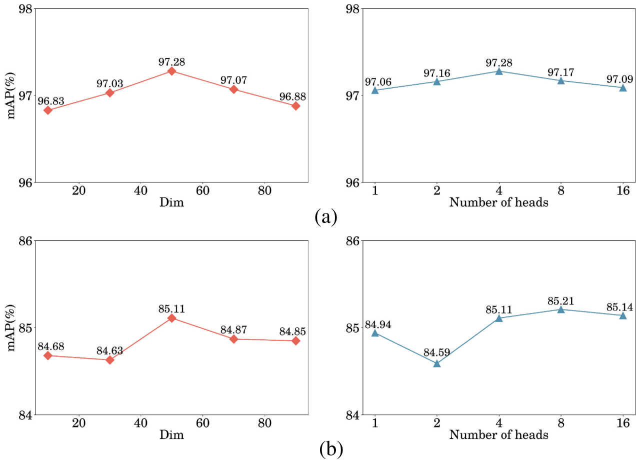
1) Ablation Study: To evaluate the effectiveness of each component of our ML-HKG, we conduct a series of ablation experiments shown in Table V. Compared with the backbone network, the visual–semantic FA module exhibits a certain improvement in EF_1 and LF_1. In addition, the semantic-object knowledge interaction (KI) enables a performance improvement by updating the imagewise semantic correlation between scenes and objects. Moreover, the scene-aware feature enhancement (SFE) module drastically enhances the model's performance to 96.83% and 84.92% of mAP on the UCM and AID datasets, indicating the efficacy of scene-level perception in benefiting multilabel classification as auxiliary knowledge. With each component's effectiveness evaluated and verified, our ML-HKG has achieved significant improvement with an



(a)

(b)

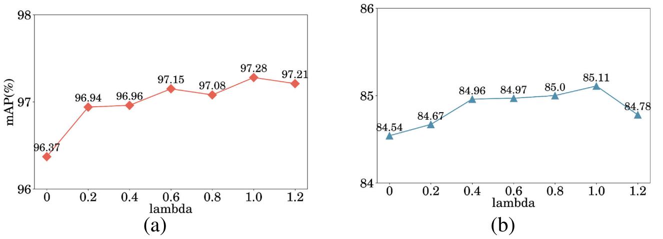
Fig. 7. Performance evaluation of different knowledge embedding strategies. (a) UCM. (b) AID.



(a)

(b)

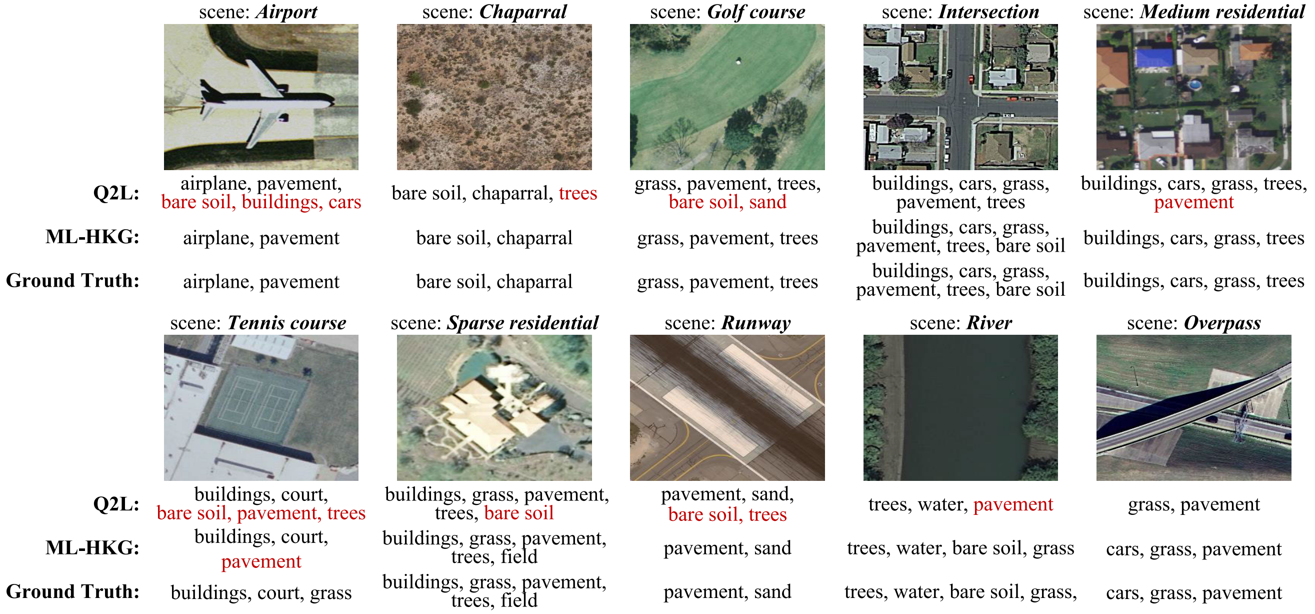
Fig. 8. Performance evaluation of different hyperparameters. (a) Hyperparameter analysis on the UCM dataset. (b) Hyperparameter analysis on the AID dataset.



(a)

(b)

Fig. 9. Performance evaluation of different values of balance weight λ. (a) Effect of λ on UCM. (b) Effect of λ on AID.



	scene: ***Airport***	scene: ***Chaparral***	scene: ***Golf course***	scene: ***Intersection***	scene: ***Medium residential***
Q2L:	airplane, pavement, bare soil, buildings, cars	bare soil, chaparral, trees	grass, pavement, trees, bare soil, sand	buildings, cars, grass, pavement, trees	buildings, cars, grass, trees, pavement
ML-HKG:	airplane, pavement	bare soil, chaparral	grass, pavement, trees	buildings, cars, grass, pavement, trees, bare soil	buildings, cars, grass, trees
Ground Truth:	airplane, pavement	bare soil, chaparral	grass, pavement, trees	buildings, cars, grass, pavement, trees, bare soil	buildings, cars, grass, trees

	scene: ***Tennis course***	scene: ***Sparse residential***	scene: ***Runway***	scene: ***River***	scene: ***Overpass***
Q2L:	buildings, court, bare soil, pavement, trees	buildings, grass, pavement, trees, bare soil	pavement, sand, bare soil, trees	trees, water, pavement	grass, pavement
ML-HKG:	buildings, court, pavement	buildings, grass, pavement, trees, field	pavement, sand	trees, water, bare soil, grass	cars, grass, pavement
Ground Truth:	buildings, court, grass	buildings, grass, pavement, trees, field	pavement, sand	trees, water, bare soil, grass,	cars, grass, pavement

Fig. 10. Comparison of prediction results between Q2L and our ML-HKG on the UCM dataset, where the incorrectly predicted labels are highlighted in red color.

increase in mAP of 1.28% on the UCM dataset and 1.41% on the AID dataset, demonstrating its superiority in MLIC tasks.

2) Effect of HKG Embedding: To analyze the effectiveness of our HKG embedding method, we replace it with the pretrained text encoders from foundation Models, i.e., CLIP [55] and Bert [56] to assess their performance under the same experimental conditions. Moreover, we additionally devise a random way and one-hot way to learn label correlations. Specifically, the random-based method treats both the scene-level and object-level label embeddings as random learnable parameters. As for the one-hot setting, we encode those labels into two independent groups of one-hot vectors. Experimental results are presented in Fig. 7, from which it can be found that the random-based embedding approach attains the worst overall performance on both UCM and AID datasets. It is also noted that the CLIP-based embedding method obtains a relatively poor result, which may be affected by the limitation of single-label-based prompts in representing multilabel correlations. In particular, we exclude textual knowledge in our method and evaluate the performance under the visual-knowledge-only (VK-only) condition, which achieves 96.9% of mAP on the UCM and 84.8% on the AID dataset, attaining the suboptimal result. Also, with the incorporation of both visual and textual knowledge, our HKG embedding approach is further improved on all metrics. In conclusion, our proposed HKG shows a significant performance improvement in multilabel classification, demonstrated by its superiority in exploring and representing the relevance not only across multilabel objects but also between object-level and scene-level labels of RS images.

3) Effect of Hyperparameters: Furthermore, to evaluate the performance of our ML-HKG under different hyperparameter settings, we analyze the sensitivity of the dimension of knowledge embeddings d_k and the effect of different numbers of heads, and the results are presented in Fig. 8. The left parts of Fig. 8(a) and (b) reflect the effect of different knowledge dimensions. As shown, the performance slightly rises with the increase of d_k and attains the optimal result on both the UCM and AID datasets at the dimension of 50. Based on that observation, the embedding dimension of our HKG is fixed to 50. Then, the right part of both subfigures shows the performance results with different numbers of heads in the multihead attention mechanism of our model, which exhibits a tiny variation as the numbers of multihead differ, which also reveals the robustness of our method. Considering the balance between model performance and efficiency, we set the number of multiple heads to 4 throughout the process.

4) Effect of Balance Weight λ: In addition, to explore the effect of different values of λ on multilabel classification performance, we also make further experiments on different values of λ, and the experimental results are shown in Fig. 9. As can be seen from the figure, the experimental results slightly fluctuate when the values of λ are taken to different values, and our proposed ML-HKG method achieves 97.28% of mAP and 85.11% of mAP when λ is set to 1, reaching the optimal results the best results. Also, it is noted that when λ is set to 0, which means that the scene-aware cross-entropy loss is not incorporated during the training process, our ML-HKG attains the worst performance, with 96.37% mAP on UCM and 84.54% mAP on AID. From the above observation, we can conclude that our model can achieve an optimal performance with not only the incorporation of scene-level supervision but also a suitable balance between multilabel loss and scene-aware loss.

G. Visualization and Analysis

1) Visual Analysis of Label Prediction: To qualitatively analyze the effectiveness of our ML-HKG, we select several RS images of representative scenes, including airport, chaparral, overpass, intersection, and river, and compare the prediction results of our model with a state-of-the-art visual

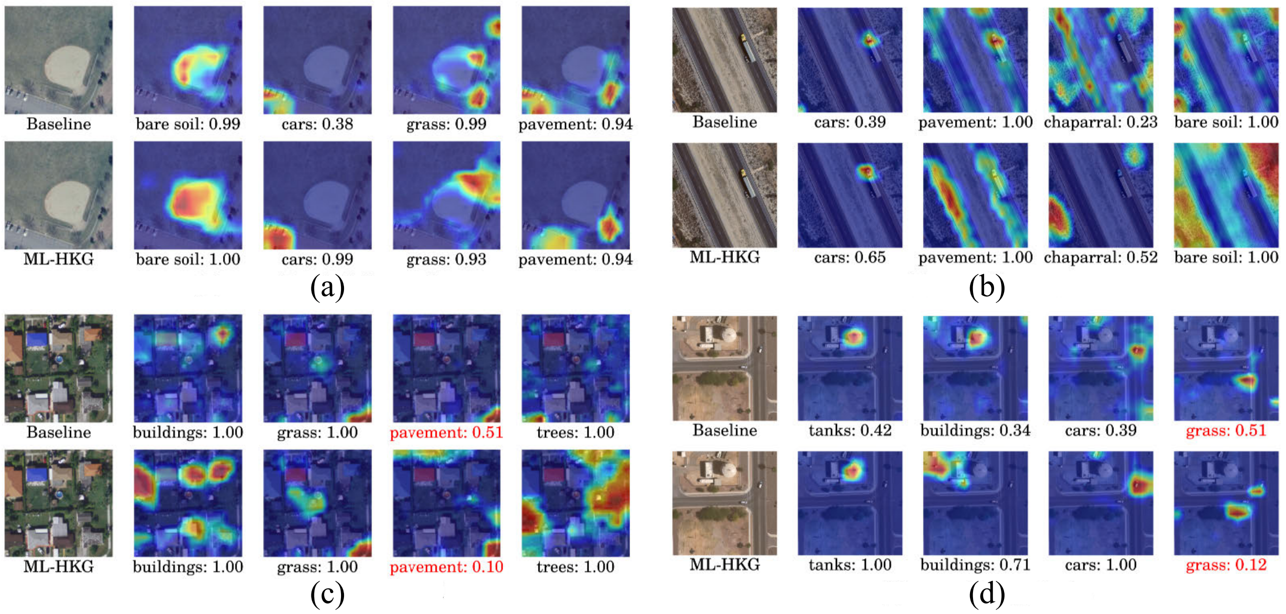


Fig. 11. Visualization analysis of baseline and our ML-HKG. For each RS scene image, we present activation maps of four object labels for evaluation, and the label not presented in the image is highlighted in red. (a) Scene: baseball diamond. (b) Scene: freeway. (c) Scene: medium residential. (d) Scene: storage tanks.

model Q2L [13] on the UCM dataset, which is shown in Fig. 10. It can be found that by learning the hierarchical label correlations, our model can make more accurate prediction results. For instance, ML-HKG can successfully predict the small-scale object label cars in the scene image of overpass, which is neglected by Q2L. Moreover, provided that the scene to which an image belongs is accurately predicted, our model can further distinguish between some confusing objects with similar visual characteristics. For instance, by successfully identifying the scene chaparral, ML-HKG can differentiate between chaparral and trees in spite of their visual similarity, thus obtaining a better multilabel classification result. In addition, our ML-HKG can also make a more reliable and rigorous multilabel prediction by introducing the hierarchical knowledge and further fusing it with visual features.

2) Visualization of Object-Specific Feature Maps: To further verify the effectiveness of our proposed ML-HKG framework, we additionally utilize Grad-CAM [57] to exhibit the object-specific activation heatmaps of the baseline model and our ML-HKG on different RS scenes, which are shown in Fig. 11. Benefiting from the scene-aware feature enhancement module, our approach can better capture the dominant objects that make up the specific RS scene during scene-level perception. For example, by perceiving the image scene as freeway, our ML-HKG can accurately identify the key objects car and pavement in Fig. 11(b), while the former object is ignored by the baseline model. Moreover, after capturing the discriminations of dominant objects, our ML-HKG can further focus on other objects with similar appearances. As shown in Fig. 11(d), after recognizing the scene storage tanks and its dominant object tanks, our model can better distinguish the visual difference between buildings and tanks, thus making better localization and identification results.

V. Conclusion

In this article, we consider the multilabel classification of RS images as a global-to-local perceiving process and, thus, propose a novel HKG-based framework. By incorporating visual co-occurrence and human textual knowledge, the HKG is constructed for comprehensively mining label correlations between scenes and multilabel objects within RS imagery. Furthermore, to obtain enhanced representations of objects for multilabel classification, the hierarchical knowledge is encoded into scene-level and object-level embeddings to guide a hierarchical visual–semantic interaction in the subsequent cross-modality decoder. Experimental results on commonly used datasets demonstrate the superiority of our ML-HKG.

Our approach demonstrates significant potential in exploring the scene-to-object recognition process that is consistent with the top–down process in human visual perceptual learning. It also enhances the understanding of complex RS environments, such as urban areas or agricultural fields. However, a current limitation is the reliance on knowledge from external sources, which may not always provide accurate or domain-specific information for RS. In the future, we aim to build a more comprehensive RS knowledge system with a larger volume of domain-specific knowledge to depict corrections in RS images with more accuracy. Furthermore, we will explore more effective knowledge representation strategies and visual–semantic interaction processes inspired by cognitive functions of the human brain, facilitating more efficient applications of multimodal data in the field of RS.

References

[1] Y. Yuan, J. Fang, X. Lu, and Y. Feng, “Remote sensing image scene classification using rearranged local features,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1779–1792, Mar. 2019.

[2] X. Zhang et al., “Remote sensing object detection meets deep learning: A metareview of challenges and advances,” *IEEE Geosci. Remote Sens. Mag.*, vol. 11, no. 4, pp. 8–44, Dec. 2023.

[3] X. Tang, M. Li, J. Ma, X. Zhang, F. Liu, and L. Jiao, "EMTCAL: Efficient multiscale transformer and cross-level attention learning for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5626915.

[4] Q. Bi, K. Qin, H. Zhang, and G.-S. Xia, "Local semantic enhanced ConvNet for aerial scene recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 6498–6511, 2021.

[5] X. Zhang et al., "Spectral–spatial distribution consistent network based on meta-learning for cross-domain hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5520915.

[6] J. Peng, L. Li, and Y. Y. Tang, "Maximum likelihood estimation-based joint sparse representation for the classification of hyperspectral remote sensing images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 6, pp. 1790–1802, Jun. 2019.

[7] Q. Bi, K. Qin, Z. Li, H. Zhang, K. Xu, and G.-S. Xia, "A multiple-instance densely-connected ConvNet for aerial scene classification," *IEEE Trans. Image Process.*, vol. 29, pp. 4911–4926, 2020.

[8] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 3, pp. 1349–1362, Mar. 2016.

[9] X. Zhang, X. Fan, G. Wang, P. Chen, X. Tang, and L. Jiao, "MFGNet: Multibranch feature generation networks for few-shot remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5609613.

[10] B. Gao and H. Zhou, "Learning to discover multi-class attentional regions for multi-label image recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 5920–5932, 2021.

[11] T. Ridnik, G. Sharir, A. Ben-Cohen, E. Ben-Baruch, and A. Noy, "ML-Decoder: Scalable and versatile classification head," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 32–41.

[12] R. You, Z. Guo, L. Cui, X. Long, Y. Bao, and S. Wen, "Cross-modality attention with semantic graph embedding for multi-label classification," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12709–12716.

[13] S. Liu, L. Zhang, X. Yang, H. Su, and J. Zhu, "Query2Label: A simple transformer way to multi-label classification," 2021, *arXiv:2107.10834*.

[14] W. Zhou, Z. Xia, P. Dou, T. Su, and H. Hu, "Aligning image semantics and label concepts for image multi-label classification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 2, pp. 1–23, May 2023.

[15] Z.-M. Chen, X.-S. Wei, X. Jin, and Y. Guo, "Multi-label image recognition with joint class-aware map disentangling and label correlation embedding," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 622–627.

[16] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5177–5186.

[17] T. Chen, M. Xu, X. Hui, H. Wu, and L. Lin, "Learning semantic-specific graph representation for multi-label image recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 522–531.

[18] Q. Bi, B. Zhou, K. Qin, Q. Ye, and G.-S. Xia, "All grains, one scheme (AGOS): Learning multigrain instance representation for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5629217.

[19] M. Ahissar and S. Hochstein, "The reverse hierarchy theory of visual perceptual learning," *Trends Cognit. Sci.*, vol. 8, no. 10, pp. 457–464, Oct. 2004.

[20] Q. Wang, Z. Mao, B. Wang, and L. Guo, "Knowledge graph embedding: A survey of approaches and applications," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 12, pp. 2724–2743, Dec. 2017.

[21] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1576–1585.

[22] R. Speer, J. Chin, and C. Havasi, "ConceptNet 5.5: An open multilingual graph of general knowledge," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, no. 1, pp. 1–8.

[23] Y. Wei et al., "HCP: A flexible CNN framework for multi-label image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1901–1907, Jun. 2015.

[24] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Aug. 2014.

[25] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2285–2294.

[26] Y. Wu, H. Liu, S. Feng, Y. Jin, G. Lyu, and Z. Wu, "GM-MLIC: Graph matching based multi-label image classification," 2021, *arXiv:2104.14762*.

[27] X. Zhu, J. Liu, W. Liu, J. Ge, B. Liu, and J. Cao, "Scene-aware label graph learning for multi-label image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 1473–1482.

[28] A. Zeggada, F. Melgani, and Y. Bazi, "A deep learning approach to UAV image multilabeling," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 694–698, May 2017.

[29] Y. Hua, L. Mou, and X. X. Zhu, "Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 149, pp. 188–199, Mar. 2019.

[30] Q. Tan, Y. Liu, X. Chen, and G. Yu, "Multi-label classification based on low rank representation for image annotati," *Remote Sens.*, vol. 9, no. 2, p. 109, Jan. 2017.

[31] X. Tan, Z. Xiao, J. Zhu, Q. Wan, K. Wang, and D. Li, "Transformer-driven semantic relation inference for multilabel classification of high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 1884–1901, 2022.

[32] B. Ma et al., "Label-driven graph convolutional network for multilabel remote sensing image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 2245–2255, 2024.

[33] T. Song, S. Bai, F. Yang, C. Gao, H. Chen, and J. Li, "Exploring hybrid contrastive learning and scene-to-label information for multilabel remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5631214.

[34] A. Singhal et al., "Introducing the knowledge graph: Things, not strings," *Off. Google Blog*, vol. 5, no. 16, p. 3, 2012.

[35] Z. Chen, Y. Wang, B. Zhao, J. Cheng, X. Zhao, and Z. Duan, "Knowledge graph completion: A review," *IEEE Access*, vol. 8, pp. 192435–192456, 2020.

[36] K. Marino, R. Salakhutdinov, and A. Gupta, "The more you know: Using knowledge graphs for image classification," 2016, *arXiv:1612.04844*.

[37] D. Lin, J. Lin, L. Zhao, Z. J. Wang, and Z. Chen, "Multilabel aerial image classification with a concept attention graph neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2021.

[38] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[39] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 1–9.

[40] Y. Lin, X. Han, R. Xie, Z. Liu, and M. Sun, "Knowledge representation learning: A quantitative review," 2018, *arXiv:1812.10901*.

[41] Y. Li, D. Kong, Y. Zhang, Y. Tan, and L. Chen, "Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 179, pp. 145–158, Sep. 2021.

[42] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Proc. AAAI Conf. Artif. Intell.*, 2015, vol. 29, no. 1, pp. 1–7.

[43] M. Jia et al., "Visual prompt tuning," in *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*. Cham, Switzerland: Springer, 2022, pp. 709–727.

[44] W. Chen, C. Si, Z. Zhang, L. Wang, Z. Wang, and T. Tan, "Semantic prompt for few-shot image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 23581–23591.

[45] W. Wang, Y. Sun, W. Li, and Y. Yang, "TransHP: Image classification with hierarchical prompting," 2023, *arXiv:2304.06385*.

[46] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[48] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. 18th SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, Nov. 2010, pp. 270–279.

[49] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, "Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1144–1158, Feb. 2018.

[50] Y. Hua, L. Mou, and X. X. Zhu, “Relation network for multilabel aerial image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4558–4572, Jul. 2020.

[51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[52] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, “Randaugment: Practical automated data augmentation with a reduced search space,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 702–703.

[53] T. DeVries and G. W. Taylor, “Improved regularization of convolutional neural networks with cutout,” 2017, *arXiv:1708.04552*.

[54] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” 2017, *arXiv:1711.05101*.

[55] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[56] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2018, *arXiv:1810.04805*.

[57] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

Xiangrong Zhang (Senior Member, IEEE) received the B.S. and M.S. degrees in computer application technology from the School of Computer Science, Xidian University, Xi’an, China, in 1999 and 2003, respectively, and the Ph.D. degree in pattern recognition and intelligent system from the School of Electronic Engineering, Xidian University, in 2006.

From January 2015 to March 2016, she was a Visiting Scientist with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. She is currently a Professor with the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education, Xidian University. Her research interests include pattern recognition, machine learning, and remote sensing image analysis and understanding.

Wenhao Hong received the B.S. degree in engineering from Xidian University, Xi’an, China, in 2022, where he is currently pursuing the M.S. degree in computer science with the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education.

His research interests include deep learning and image classification.

Zhenyu Li received the B.S. and M.S. degrees in intelligent science and technology from Xidian University, Xi’an, China, in 2019 and 2022, respectively, where he is currently pursuing the Ph.D. degree with the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education.

His research interests include digital image processing and federated learning.

Xina Cheng (Member, IEEE) received the B.E. degree from the School of Optoelectronics, Beijing Institute of Technology, Beijing, China, in 2014, and the M.E. and Ph.D. degrees from the Graduate School of Information, Production and Systems, Waseda University, Tokyo, Japan, in 2015 and 2018, respectively.

She is currently a Lecturer with Xidian University, Xian, China. Her research interests include sports analysis and computer vision.

Xu Tang (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electronic circuit and system from Xidian University, Xi’an, China, in 2007, 2010, and 2017, respectively.

From 2015 to 2016, he was a Joint Ph.D. along with Prof. W. J. Emery at the University of Colorado at Boulder, Boulder, CO, USA. He is currently an Associate Professor with the Key Laboratory of Intelligent Perception and Image Understanding, Ministry of Education, Xidian University. His current research interests include remote sensing image content-based retrieval and reranking, hyperspectral image processing, remote sensing scene classification, and object detection.

Huiyu Zhou received the B.Eng. degree in radio technology from Huazhong University of Science and Technology, Wuhan, China, in 1990, the M.Sc. degree in biomedical engineering from the University of Dundee, Dundee, U.K., in 2002, and the Ph.D. degree in computer vision from Heriot-Watt University, Edinburgh, U.K., in 2006.

He is currently a Full Professor with the School of Computing and Mathematical Sciences, University of Leicester, Leicester, U.K. He has authored or co-authored over 380 peer-reviewed articles in the field.

Licheng Jiao (Fellow, IEEE) received the B.S. degree from Shanghai Jiao Tong University, Shanghai, China, in 1982, and the M.S. and Ph.D. degrees from Xi’an Jiaotong University, Xi’an, China, in 1984 and 1990, respectively.

Since 1992, he has been a Professor with the School of Artificial Intelligence, Xidian University, Xi’an, where he is the Director of the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China. He is in charge of about 40 important scientific research projects and has authored over 20 monographs and 100 papers in international journals and conferences. His research interests include image processing, natural computation, machine learning, and intelligent information processing.