



Label-Guided Cross-Modal Attention Network for Multi-Label Aerial Image Classification

Ying Chen , Ding Zhang, Tao Han , *Member, IEEE*, Xiaoliang Meng, Mianxin Gao, Teng Wang

Abstract—Multi-label aerial image classification is a fundamental yet complex task in remote sensing interpretation that aims to identify multiple labels in a single image. In this letter, we propose a Label-Guided Cross-Modal Attention (L-GCMA) network, which first introduces a novel approach to enriching the semantic information of labels and utilizes the multi-head attention module to extract diverse features. The proposed method consists of two components before the cross-modal attention. Firstly, the visual features of the image are obtained using a transformer encoder. Additionally, to capture the rich semantic relationship of the scene, we design a Label-Sentence Mapping Attention (L-SMA) module. This module performs word embedding encoding on the labels and applies BERT encoding on the sentences, followed by multi-head attention to extract comprehensive inter- and intra-class relationships for the labels, specifically obtaining label-scene text features. Subsequently, by treating the text features as a query, the visual features and text features are combined using cross-modal attention. This progressive integration narrows the semantic gap between vision and text, facilitating accurate label recognition. Our proposed L-GCMA consistently achieves state-of-the-art performance on the multi-label aerial image classification task, as demonstrated by extensive experiments on two visual benchmarks - the UCM multi-label dataset and the AID multi-label dataset.

Index Terms—BERT, multi-label image classification, multi-head attention, cross-modal, aerial image.

I. INTRODUCTION

AS the application of high-resolution aerial images becomes more extensive, the classification of aerial images is gaining increasing attention. However, due to the presence of multiple classes of geographic objects in aerial images, multi-label classification (MLC) becomes both applicable and practical.

Manuscript received April 19, *; revised September 17, *. This work was supported in part by Shanghai Pujiang Program 22PJ1423400 and in part by the Surveying and Mapping Institute, Lands and Resource Department of Guangdong Province Program JDZ22035. (Corresponding author: Teng Wang. First affiliation: China Telecom Research Institute.)

Ying Chen is with the China Telecom Research Institute, Shanghai 200120, China, and also with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China(email: chen323@chinatelecom.cn).

Tao Han is with the China Telecom Research Institute, Shanghai 200120, China (email: hant2@chinatelecom.cn).

Ding Zhang and Xiaoliang Meng are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (email: frankzhang@whu.edu.cn; xmeng@whu.edu.cn).

Mianxin Gao and Teng Wang are with the Surveying and Mapping Institute, Lands and Resource Department of Guangdong Province, Guangzhou 510663, China, Key Laboratory of Natural Resources Monitoring in Tropical and Subtropical Area of South China, Ministry of Natural Resources, Guangzhou 510663, China, and also with the Guangdong Science and Technology Collaborative Innovation Center for Natural Resources, Guangzhou 510663, China (e-mail: gmxgd@163.com; wangteng43@hotmail.com).

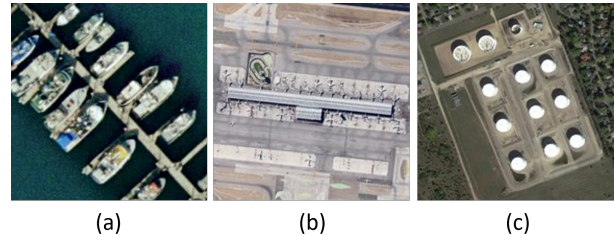


Fig. 1: Example high-resolution aerial images with their scene labels and multiple object labels. (a) harbor: *dock, ship, water*. (b) airport: *airplane, bare-soil, buildings, pavement*. (c) tanks: *tanks, bare-soil, buildings, cars, grass, trees*.

MLC aims to improve the recognition of a diverse range of object labels within an image. These object-level labels have a potential relationship with the entire scene depicted in the image. To illustrate this, refer to Fig. 1, which provides an example of multi-label aerial images. These images include object labels that not only describe individual objects but also imply their interplay with the overall scene context. Fig. 1 (a) shows a harbor scene with ships moored at the docks and floating on the water. This relationship between the scene and the objects depicted aligns with our prior knowledge. Therefore, to achieve improved efficacy in MLC for aerial images, it is necessary to comprehensively utilize the relationships between labels and scenes.

In recent years, there has been an increasing interest in studying the correlations between labels in aerial images. Various approaches have been developed to leverage the feature extraction capabilities of CNNs [1], [7], model relationships between adjacent labels using RNNs [15], [10], and utilize graph connections to learn potential relationships through region adjacency graphs [3], [12]. However, these methods have limitations in effectively capturing distinct object features and aligning semantic features of the image with the label concept. Furthermore, they do not fully exploit the information contained within the labels.

To address the mentioned issues, we propose a novel network called Label-Guided Cross-Modal Attention (L-GCMA). Taking inspiration from Query2Label [13] and TSFormer [18], we begin by transforming the input image into patches, which facilitates the extraction of visual features. Additionally, we introduce a module called Label-Sentence Mapping Attention (L-SMA) to effectively capture the relationships between labels. This module combines prior knowledge of aerial images with the powerful contextual semantic association capabilities of BERT [5]. Finally, we utilize multi-head attention to merge

the visual and textual modalities, allowing for comprehensive learning of label information and establishing its correspondence to relevant regions within the image.

In this letter, we have made the following contributions:

- 1) We have proposed a Label-Guided Cross-Modal Attention (L-GCMA) network that uses a cross-modal attention mechanism within an end-to-end framework to integrate text and visual information.
- 2) We have designed a Label-Sentence Mapping Attention (L-SMA) module, which constructs sentence prompts of labels based on prior knowledge of the data and thoroughly explores the relationship between labels and scenes through multi-head attention.
- 3) To our knowledge, this is the first time we have utilized cross-modal attention to connect the labels and aerial images. Extensive experimental results demonstrate that our L-GCMA achieves state-of-the-art performance in multi-label aerial image classification tasks.

II. METHODOLOGY

Given an input aerial image, represented as a tensor $x \in \mathbb{R}^{H_0 \times W_0 \times 3}$, Multi-Label Classification (MLC) aims to predict the presence of various categories of interest. These categories can be object classes like *airplane*, *cars*, and *ships*, or scene categories like *grass* and *bare-soil*. Let there be a total of C categories, and we denote the corresponding label of x as $y = [y_1, \dots, y_C]$, where $y_c \in \{0, 1\}$ is a binary indicator. If the c -th category label is present in the image x , then $y_c = 1$, otherwise $y_c = 0$. Our model takes x as input and predicts the probabilities of the presence of each category, represented as $p = [p_1, \dots, p_C]$, where $p_c \in [0, 1]$.

Fig. 2 shows our Label-Guided Cross-Modal Attention (L-GCMA) network for MLC of aerial images. The input image goes through the patch embedding layer to get visual embeddings, and then through the transformer encoder to learn visual representations. The label set goes through the word embedding layer and the label-sentence mapping to get BERT embeddings of sentence prompts. These embeddings are used in a multi-head attention mechanism to learn representations for the label's textual features. The visual and textual features are merged using cross-modal attention to aggregate context-aware patch features. Finally, a set of binary classifiers uses the learned label-specific visual representations to predict the presence of corresponding labels in the image.

A. Vision Feature Extracting

For visual feature extraction, we use the standard vision transformer [6]. Aerial image x , undergoes a transformation into a sequence of flattened 2D patches, serving as the input for the transformer. Following the patch embedding layer, the [class] token and positional embeddings are successively incorporated. The transformer encoder, with L layers, processes the input sequence, culminating in visual features $\mathcal{F}_{vision} \in \mathbb{R}^{HW \times D}$, where HW represents the height and weight of the feature map, and D is the vector size of transformer, which remains constant across all its layers.

B. Label-Sentence Mapping Attention Block

To obtain sufficient information on the relationship between labels and their corresponding scenes, a co-occurrence matrix was constructed based on the co-occurrence relationship between labels and scenes, as shown in Fig. 3.

Labels and scenes with high correlation coefficients are more likely to appear together. Based on the co-occurrence matrix, we designed sentence prompts for labels with prior knowledge, as shown in Table I.

TABLE I: A prior prompt of multi-label datasets of aerial images.

Label	Prompt
airplane	An aerial photo of airplane.
bare-soil	Baseball diamond and chaparral contains bare soil.
buildings	Dense residential, medium residential, and sparse residential contain buildings.
cars	Freeway, intersection, medium residential, and parking lot contain cars.
chaparral	An aerial photo of chaparral.
court	An aerial photo of court.
dock	Harbor contains dock.
field	Agricultural contains field.
grass	Golf course contains grass.
mobile-home	Mobile home park contains mobile-home.
pavement	Airplane, freeway, intersection, mobile home park, overpass, parking lot, and runway contain pavement.
sand	Beach contains sand.
sea	Beach contains sea.
ship	Harbor contains ship.
tanks	Storage tanks contains tanks.
trees	Forest contains trees.
water	Harbor and river contain water.

Analyzing the co-occurrence matrix helps establish the connection between labels and scenes. The design prompts $\{\text{An aerial photo of [label]}\}$ when there is no clear co-occurrence relationship for a label, and $\{[\text{Scene}] \text{ contains [label]}\}$ when a label has a strong co-occurrence relationship. The prompt statement begins with the [CLS] token at the beginning of the sentence and [SEP] token at the end of the sentence, like $\{[\text{CLS}], \text{'Beach'}, \text{'contains'}, \text{'sea'}, [\text{SEP}]]\}$, which is input to the BERT pre-trained model to obtain the sentence text feature $\mathcal{F}_{sentence} \in \mathbb{R}^{C \times D}$.

We have developed a label-sentence attention module that aims to explore the relationships between class labels. This module employs the word vector feature \mathcal{F}_{word} obtained via word embedding as a *query* and the sentence vector feature $\mathcal{F}_{sentence}$ of the label to obtain a complete representation of the label, i.e., $\mathcal{F}_{text} \in \mathbb{R}^{C \times D}$. The procedure is illustrated by the dashed line on the right-hand side of Fig. 2. Concretely, label-sentence attention takes the sentence embedding feature $\mathcal{F}_{sentence}$ in sentence stream and the word embedding feature \mathcal{F}_{word} in word stream as inputs, with the former serving as *query* and the latter as *key* and *value*:

$$\begin{aligned}
 Q_{word} &= \mathcal{F}_{word} W_{word-Q}, \\
 K_{sentence} &= \mathcal{F}_{sentence} W_{sentence-K}, \\
 V_{sentence} &= \mathcal{F}_{sentence} W_{sentence-V},
 \end{aligned} \tag{1}$$

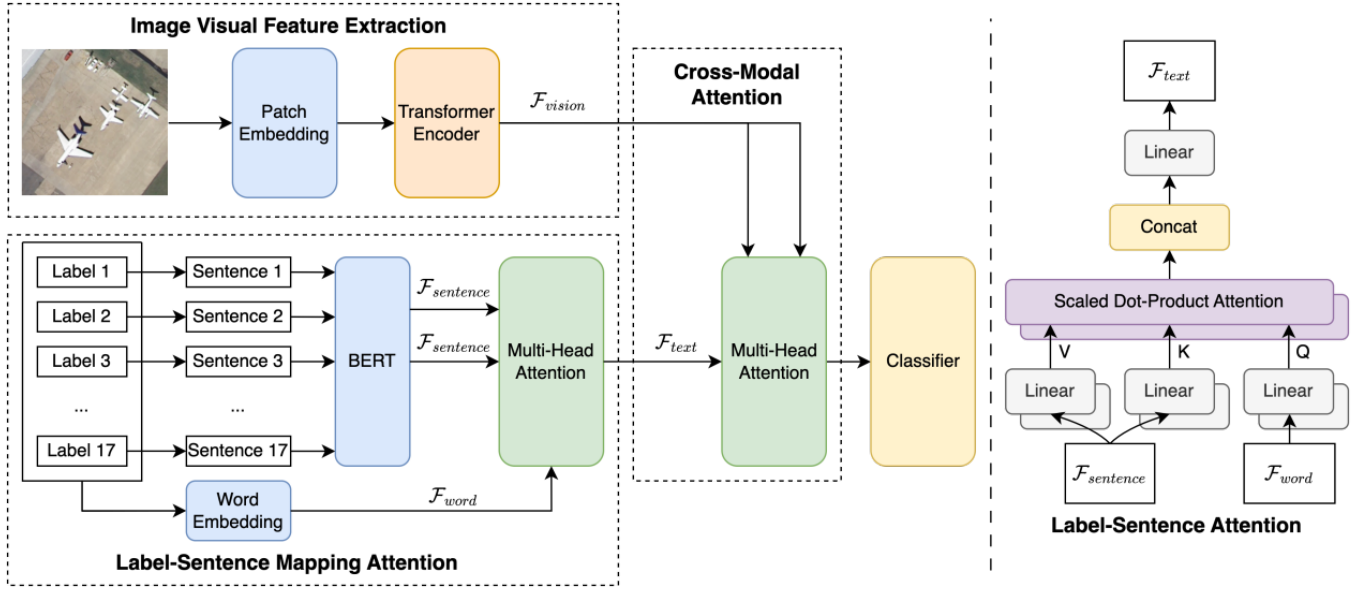


Fig. 2: The overall framework of the proposed L-GCMA, which consists of image visual feature extraction for capturing visual features, label-sentence mapping attention for learning label representations, and cross-modal attention for integrating vision and text features. The left part illustrates the data pipeline in L-GCMA, and the right part focuses on the details of the label-sentence attention.

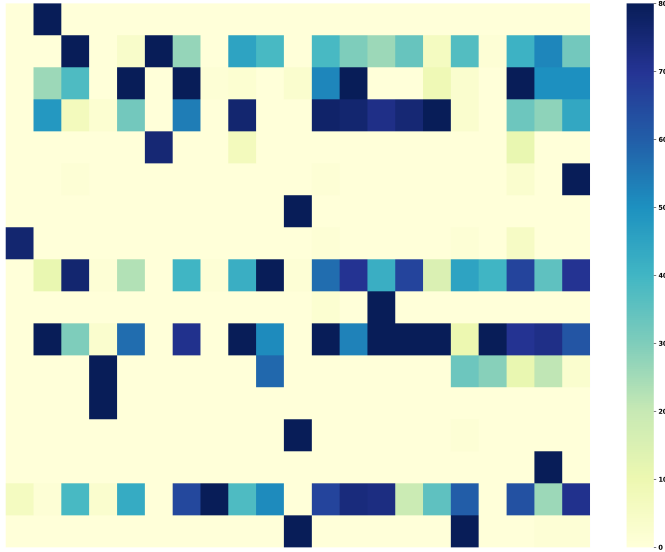


Fig. 3: Label and scene correlation matrix. The horizontal axis represents the labels and the vertical axis represents the scenes.

where $W_{word-Q} \in \mathbb{R}^{D \times D}$, $W_{sentence-K} \in \mathbb{R}^{D \times D}$ and $W_{sentence-V} \in \mathbb{R}^{D \times D}$ are parameter matrices to be learned. Then the label's textual feature is obtained by:

$$\mathcal{F}_{text} = \text{MultiHead}(Q_{word}, K_{sentence}, V_{sentence}). \quad (2)$$

C. Text-Vision cross-modal attention

We utilize label embeddings as queries represented by \mathcal{F}_{text} and execute cross-attention to gather class-specific features from the visual feature \mathcal{F}_{vision} . Our approach employs the same multi-head attention structure as L-SMAB, with the label

embedding serving as the query, while the key and value correspond to visual features, as follows:

$$\begin{aligned} Q_{text} &= \mathcal{F}_{text} W_{text-Q}, \\ K_{vision} &= \mathcal{F}_{vision} W_{vision-K}, \\ V_{vision} &= \mathcal{F}_{vision} W_{vision-V}, \end{aligned} \quad (3)$$

where $W_{text-Q} \in \mathbb{R}^{D \times D}$, $W_{vision-K} \in \mathbb{R}^{D \times D}$ and $W_{vision-V} \in \mathbb{R}^{D \times D}$ are parameter matrices to be learned. Finally, the label-specific visual representations $\mathcal{F} \in \mathbb{R}^{C \times D}$ is obtained by:

$$\mathcal{F} = \text{MultiHead}(Q_{text}, K_{vision}, V_{vision}). \quad (4)$$

D. Loss Function

For the task of multi-label aerial image classification, we consider the classification of each label as a binary classification problem. This means that we can leverage label-specific visual features \mathcal{F} obtained through cross-modal fusion to calculate the probability of prediction. The formulation is as follows:

$$p = \text{Sigmoid}(f(\mathcal{F} \odot W_C)), \quad (5)$$

where $p \in \mathbb{R}^C$ is the corresponding probability vector, $f(\cdot)$ represents the operation of summing the rows of a matrix, \odot denotes element-wise multiplication and $W_C \in \mathbb{R}^{D \times 1}$ is a learnable parameter matrix. The binary cross entropy is used as the loss function, which is formulated as follows:

$$\mathcal{L} = - \sum_{c=1}^C (y_c \log p_c + (1 - y_c) \log (1 - p_c)). \quad (6)$$

III. EXPERIMENTS

A. Datasets

UCM multi-label dataset consists of 2100 RGB aerial images [2]. This dataset includes 17 object classes: *airplane*, *sand*, *pavement*, *building*, *car*, *chaparral*, *court*, *tree*, *dock*, *tank*, *water*, *grass*, *mobile-home*, *ship*, *bare-soil*, *sea*, and *field*. Each image has a resolution of 256×256 pixels and a spatial resolution of 0.3 meters. Each image has one or more labels (up to 7). On average, each image contains 3.3 object-level labels.

AID multi-label dataset is a compilation of 3000 aerial images [9], which have been labeled in alignment with the UCM multi-label dataset. Each image boasts a resolution of 600×600 pixels and a spatial resolution ranging from 0.5 to 8 meters. On average, each image is associated with 5.5 object-level labels, with a maximum label count of 11.

B. Evaluation Metrics

To effectively compare the proposed method with others, performance evaluation metrics beyond mean average precision (mAP) are necessary. According to [4], the metrics include example-based precision (EP), recall (ER), F1-scores (EF_1), and label-based precision (LP), recall (LR), and F1-score (LF_1). These metrics are computed as follows:

$$\begin{aligned} EP &= \frac{\sum_i M_c^i}{\sum_i M_p^i}, & ER &= \frac{\sum_i M_c^i}{\sum_i M_g^i}, \\ LP &= \frac{1}{C} \sum_i \frac{M_c^i}{M_p^i}, & LR &= \frac{1}{C} \sum_i \frac{M_c^i}{M_g^i}, \\ EF_1 &= \frac{2 \times EP \times ER}{EP + ER}, & LF_1 &= \frac{2 \times LP \times LR}{LP + LR}, \end{aligned} \quad (7)$$

where M_c^i is the number of images predicted correctly for the i -th category, M_p^i is the number of images predicted for the i -th category, and M_g^i is the number of ground truth images for the i -th category.

C. Implementations Details

For extracting visual features from aerial images, we utilize the ViT-B16 [6] model of the Vision Transformer, which is pre-trained on ImageNet21k. In addition to this, we use the 768-dimensional representations generated by BERT [5] as initial label embeddings. To ensure a fair comparison with competitors, we resize all input images to 256×256 during both the training and testing phases of all experiments. The label categories of the two datasets are divided into train, validation, and test sets with a ratio of 72%, 8%, and 20% in experiments. The entire network is trained using the AdamW algorithm, with a batch size of 16 and momentums of 0.999 and 0.9. The learning rate is initialized at 1×10^{-5} , and decays by a factor of 10 when the loss plateaus. To augment data, we perform random horizontal flips, and random resized crops during the training stage. All experiments are conducted on a server equipped with 4 NVIDIA 3090 GPUs.

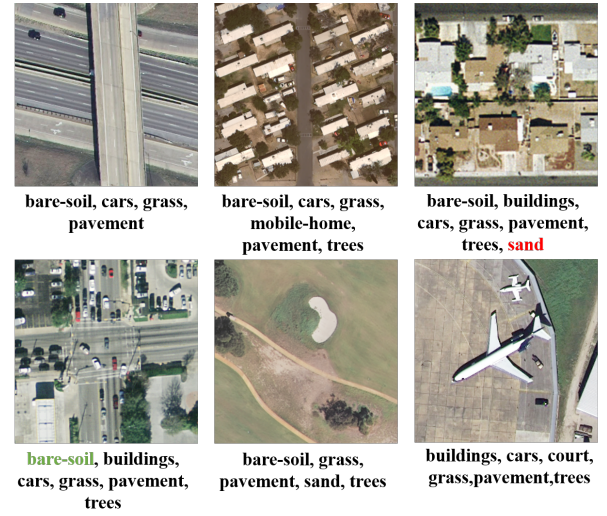


Fig. 4: Example images and predicted labels by our L-GCMA on the UCM multi-label dataset. Red predictions indicate false positives, while green predictions are false negatives.

D. Results

To demonstrate the effectiveness of our L-GCMA method, we conduct extensive experiments on two widely used and challenging multi-label aerial image classification benchmark datasets: UCM- and AID multi-label datasets. The selected models for comparison included ResNet [8], CA-BiLSTM [10], Stivaktakis [14], RBFNN [16], MLGCN [4], ML-CG [12], Zhu [17], and Huang [11].

Results on UCM multi-label dataset. Experimental results are presented in Table II. Our L-GCMA demonstrates superior performance, surpassing previous state-of-the-art methods by a significant margin. To the best of our knowledge, it is the first approach to achieve an EF_1 of over 93.76% on the UCM multi-label dataset.

TABLE II: Results on the UCM multi-label dataset (%)

Method	EP	ER	EF_1	LP	LR	LF_1
ResNet-50 [8]	80.86	81.95	81.4	88.78	78.98	83.59
CA-BiLSTM [10]	77.94	89.02	83.11	86.12	84.26	85.18
MLGCN [4]	79.86	82.10	80.96	86.42	80.83	83.53
ML-CG [12]	81.34	89.94	85.42	88.53	89.27	88.90
Zhu [17]	91.75	91.65	90.62	92.96	92.60	92.66
Huang [11]	90.54	92.98	91.74	93.73	92.75	93.23
L-GCMA	92.23	95.34	93.76	95.41	97.54	96.47

Some examples of the predicted results on the UCM multi-label dataset are shown in Figure 4. It is evident that the proposed method effectively captures the main categories of the scene. In order to showcase how effective text-vision cross-modal attention is, we have included some attention maps in Fig. 5. Our findings indicate that our model is able to locate the desired object with reasonable accuracy, particularly with smaller or medium-sized objects.

Results on AID multi-label dataset. We present a comparison of our method with other methods in Table III. Our proposed method demonstrates superior performance in terms of EF_1 and LF_1 , which are considered the most important metrics. Specifically, our L-GCMA outperforms Huang by

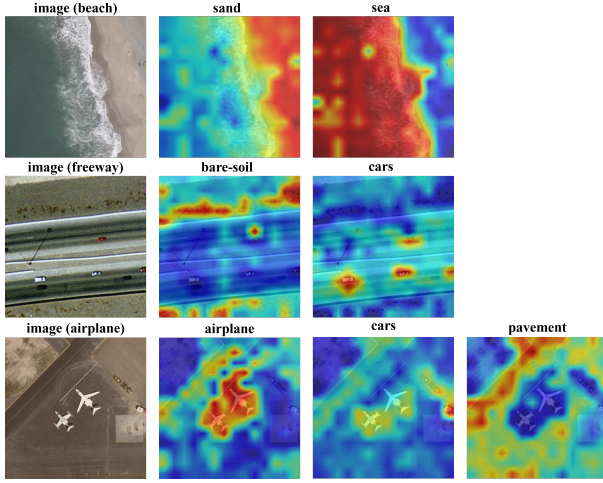


Fig. 5: Visualization of cross-attention maps on the UCM multi-label dataset. Texts above images represent the ground truth labels (query) for the scene images.

0.49%, Zhu by 4.04%, MLGCN by 2.35%, RBFNN by 4.41%, Stivaktakis by 4.14% and ResNet-50 by 9.95%.

TABLE III: Results on the AID multi-label dataset (%)

Method	EP	ER	EF ₁	LP	LR	LF ₁
ResNet-50 [8]	84.12	79.94	81.98	54.08	42.04	47.30
Stivaktakis [14]	87.69	87.92	87.79	73.40	66.45	69.74
RBFNN [16]	88.52	86.56	87.52	78.78	64.16	70.70
MLGCN [4]	89.69	89.48	89.58	78.91	75.06	76.90
Zhu [17]	89.72	88.41	87.49	80.89	74.08	76.50
Huang [11]	91.03	91.88	91.44	81.37	74.60	77.79
L-GCMA	92.19	91.68	91.93	84.39	75.60	79.76

E. Ablation Studies

To assess the efficacy of our proposed L-GCMA model and the significance of fusing image-label information, we carried out a series of ablation experiments on the UCM multi-label dataset. The outcomes of these experiments are summarized in Table IV. The findings highlight the critical role played by the L-SMAB module in capturing label-related information, leveraging prior knowledge of aerial scenes. In the context of MLC, the comprehensive incorporation of textual label information enables semantic alignment between label and image data, leading to improved accuracy in label classification.

TABLE IV: Ablation study of the proposed modules on the UCM multi-label dataset (%)

Method	mAP
ViT-B16 [6]	98.27
L-GCMA w/o BERT [5]	98.80
L-GCMA w/o L-SMAB	98.78
L-GCMA	99.10

IV. CONCLUSION

In this letter, we construct a novel label-guided cross-modal attention network named L-GCMA, for multi-label

classification of aerial images. Our method designs a label-sentence mapping attention module to obtain sufficient label feature information through prior knowledge prompts and utilize a cross-modal attention mechanism to fuse visual and textual features, alleviating the semantic gap and linking semantic information with visual information. Experimental results on two multi-label aerial image datasets demonstrate the effectiveness of our proposed method, and ablation studies confirm the importance of our core designs.

REFERENCES

- [1] Yakoub Bazi. Two-branch neural network for learning multi-label classification in uav imagery. *international geoscience and remote sensing symposium*, 2019.
- [2] Bindita Chaudhuri, Begum Demir, Subhasis Chaudhuri, and Lorenzo Bruzzone. Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method. *IEEE Transactions on Geoscience and Remote Sensing*, 2018.
- [3] Jingzhou Chen and Yuntao Qian. Hierarchical multilabel ship classification in remote sensing images using label relation graphs. *IEEE Transactions on Geoscience and Remote Sensing*, 2021.
- [4] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. *computer vision and pattern recognition*, 2019.
- [5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *north american chapter of the association for computational linguistics*, 2018.
- [6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.
- [7] Daniel Gardner and David Nichols. Multi-label classification of satellite images with deep learning. 2022.
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv: Computer Vision and Pattern Recognition*, 2015.
- [9] Yuansheng Hua, Lichao Mou, and Xiao Xiang Zhu. Label relation inference for multi-label aerial image classification. *international geoscience and remote sensing symposium*, 2019.
- [10] Yuansheng Hua, Lichao Mou, and Xiao Xiang Zhu. Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional lstm network for multi-label aerial image classification. *Isprs Journal of Photogrammetry and Remote Sensing*, 2019.
- [11] Rui Huang, Fengcai Zheng, and Wei Huang. Multilabel remote sensing image annotation with multiscale attention and label correlation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021.
- [12] Dan Lin, Jianzhe Lin, Liang Zhao, Z. Jane Wang, and Zhikui Chen. Multilabel aerial image classification with a concept attention graph neural network. *IEEE Transactions on Geoscience and Remote Sensing*, 2021.
- [13] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification. *arXiv: Computer Vision and Pattern Recognition*, 2021.
- [14] Radamanthys Stivaktakis, Grigorios Tsagkatakis, and Panagiotis Tsakalides. Deep learning for multilabel land cover scene categorization using data augmentation. *IEEE Geoscience and Remote Sensing Letters*, 2019.
- [15] Gencer Sumbul and Begum Demir. A novel multi-attention driven system for multi-label remote sensing image classification. *international geoscience and remote sensing symposium*, 2019.
- [16] Abdallah Zeggada, Farid Melgani, and Yakoub Bazi. A deep learning approach to uav image multilabeling. *IEEE Geoscience and Remote Sensing Letters*, 2017.
- [17] Panpan Zhu, Yumin Tan, Liqiang Zhang, Yuebin Wang, Jie Mei, Hao Liu, and Mengfan Wu. Deep learning for multilabel remote sensing image annotation with dual-level semantic concepts. *IEEE Transactions on Geoscience and Remote Sensing*, 2020.
- [18] Xuelin Zhu, Jiuxin Cao, Jiawei Ge, Weijia Liu, and Bo Liu. Two-stream transformer for multi-label image classification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3598–3607, 2022.