




## Article

# S-MAT: Semantic-Driven Masked Attention Transformer for Multi-Label Aerial Image Classification

Hongjun Wu <sup>1,2</sup> , Cheng Xu <sup>1,2</sup>  and Hongzhe Liu <sup>1,2,\*</sup> 

<sup>1</sup> Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing 100101, China; gwawhj@126.com (H.W.); xc-f4@163.com (C.X.)

<sup>2</sup> Institute for Brain and Cognitive Sciences, Beijing Union University, Beijing 100101, China

\* Correspondence: liuhongzhe@buu.edu.cn

**Abstract:** Multi-label aerial scene image classification is a long-standing and challenging research problem in the remote sensing field. As land cover objects usually co-exist in an aerial scene image, modeling label dependencies is a compelling approach to improve the performance. Previous methods generally directly model the label dependencies among all the categories in the target dataset. However, most of the semantic features extracted from an image are relevant to the existing objects, making the dependencies among the non-existent categories unable to be effectively evaluated. These redundant label dependencies may bring noise and further decrease the performance of classification. To solve this problem, we propose S-MAT, a Semantic-driven Masked Attention Transformer for multi-label aerial scene image classification. S-MAT adopts a Masked Attention Transformer (MAT) to capture the correlations among the label embeddings constructed by a Semantic Disentanglement Module (SDM). Moreover, the proposed masked attention in MAT can filter out the redundant dependencies and enhance the robustness of the model. As a result, the proposed method can explicitly and accurately capture the label dependencies. Therefore, our method achieves CF1s of 89.21%, 90.90%, and 88.31% on three multi-label aerial scene image classification benchmark datasets: UC-Merced Multi-label, AID Multi-label, and MLRSNet, respectively. In addition, extensive ablation studies and empirical analysis are provided to demonstrate the effectiveness of the essential components of our method under different factors.

**Keywords:** aerial scene classification; multi-label learning; redundancy removing; label correlation; semantic disentanglement



**Citation:** Wu, H.; Xu, C.; Liu, H. S-MAT: Semantic-Driven Masked Attention Transformer for Multi-Label Aerial Image Classification. *Sensors* **2022**, *22*, 5433. <https://doi.org/10.3390/s22145433>

Academic Editor: Yifan Chen

Received: 23 June 2022

Accepted: 19 July 2022

Published: 20 July 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

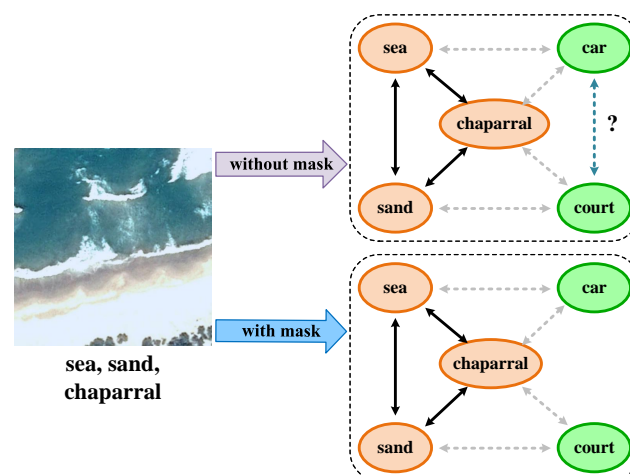
## 1. Introduction

Recently, the great progress in remote sensing technologies has provided increasing remote sensing images from satellite-borne and airborne sensors for land cover mapping and monitoring issues. Generally, since high-resolution remote sensing images depict diverse categories of land cover objects, a single label cannot accurately describe the content in the image. Therefore, compared with single-label classification [1–3], multi-label remote sensing image classification is a more practical task. Specifically, a Multi-Label Image Classification (MLIC) method is developed to assign a set of preset land cover labels to each remote sensing image. In this paper, we focus on the multi-label remote sensing image classification task in the aerial scene understanding field.

Benefiting from the great success of deep learning, deep Convolutional Neural Networks (CNNs) [4,5] and vision Transformers [6,7] are proposed to extract high-level semantic features and make incremental progress in Single-Label multi-class Image Classification (SLIC). Additionally, the field of numerical simulation and stability has achieved significant progress [8]. These advanced methods are also exploited in single-label remote sensing image classification. By treating each label in isolation, the multi-label problem can be simply addressed by using SLIC methods to predict whether each label is present or not.

However, compared to SLIC, MLIC is a more complicated task. On the one hand, in an aerial image, there are multiple land cover objects at different spatial resolutions, which are related to the size of the objects. For example, the size of a car is far less than a court, and consequently, “car” is one of the inconspicuous categories. On the other hand, since land cover objects generally co-exist in an aerial scene image, the inter-class relationship is another key for the classification. Therefore, the MLIC task considers not only accurate spatial feature extraction, but also the correlations of multiple concepts. In classical MLIC, the utilization of spatial information and inter-class correlations are both significant issues. To handle the spatial information, some works introduce regional proposal techniques [9,10], implicit spatial attention [11,12], or multi-scale features [13]. Nevertheless, these methods neglect the impact of the relationships among multiple categories. On the other, many works are proposed to model the inter-class correlations. Pioneering approaches [14–18] based on Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) learn the correlations in sequential prediction. However, the performance of the RNN-based methods is influenced by the pre-set or learned sequence. Moreover, the complex label correlations cannot be accurately represented by sequential relations. Other works [19,20] formulate the MLIC task as a structural inference problem based on probabilistic graphical models [21], while the practicality is limited by the high computational complexity. Inspired by the great success of the Graph Convolutional Network (GCN) [22] in the representation of multivariate relations, ML-GCN [23] proposes to explicitly model label correlations via GCN. Transformer [24] from the Natural Language Processing (NLP) field has achieved great success [25–27] in the Computer Vision (CV) area. Inspired by the impressive capability to model long-range dependencies, recent works [28,29] leverage Transformer to capture the label dependencies.

Previous methods have proven the effectiveness of dealing with label dependencies. In general, these methods directly model the holistic label dependencies among the generated label embeddings related to all categories. Nevertheless, only a portion of category objects is present in a single image, and the visual features extracted from the image are mostly relevant to the ground-truth labels. Consequently, the computed label dependencies among the non-existent categories are inaccurate. In this paper, these label dependencies are called redundant dependencies, which bring noise to the classification task. Specifically, as shown in Figure 1, the solid line indicates a higher connection and the dashed line indicates a lower connection between categories. According to the presence or absence of the categories in the image, some of the inter-class relationships are obvious and definite, such as the solid lines (higher connections) and dashed grey lines (lower connections). However, dependencies among non-existent categories (e.g., car and court) cannot be accurately estimated, which leads to a negative impact on classification.



**Figure 1.** Illustration of the effect of mask attention.

In short, there are mainly three challenges in multi-label remote sensing image classification tasks. Firstly, land cover objects have different scales in an image, making it more challenging to leverage spatial information. Secondly, there are usually co-occurrence dependencies among the objects of different categories in a remote sensing image. Thirdly, the redundant correlations among the nonexistent categories bring noise to multi-label remote sensing image classification.

To this end, we propose a novel and effective multi-label image classification framework, Semantic-driven Masked Attention Transformer (S-MAT), which consists of a backbone feature extractor, a Semantic Disentanglement Module (SDM), and a Masked Attention Transformer (MAT). In this paper, we first extract the high-level semantic feature of a remote sensing image with a CNN backbone. Then, the extracted feature is disentangled into a set of label embeddings in the SDM. After that, a Masked Attention Transformer (MAT) is trained to model label correlations and update the input label embeddings adaptively. Meanwhile, we exploit the masked attention in MAT to restrict the attention to the categories with higher confidence and avoid the noise from nonexistent categories. Finally, the updated label embeddings are projected to the image-level predictions, which are combined with the independent predictions generated directly from the high-level image representation to obtain the final predictions. In addition, we notice that masked attention is applied to other visual tasks, such as Mask2Former [30] for image segmentation. The attention mask in Mask2Former is class agnostic, whereas, in our work, each element  $M_{ij}$  in attention mask  $M$  corresponds to the relevance between class  $i$  and  $j$  in the attention map.

Comprehensive results on three widely used multi-label image recognition benchmarks show that our S-MAT outperforms other recent methods that model label relationships via graph convolution networks or other proposed strategies. In summary, our main contributions are as follows:

- A novel Transformer-based framework, S-MAT, namely Semantic-driven Masked Attention Transformer, is proposed. S-MAT aims to filter out the redundant dependencies and obtain more accurate label dependencies for multi-label aerial scene image classification.
- We conduct in-depth studies on the application of masked attention and propose a plug-and-play module, Masked Attention Transformer (MAT), to constrain the attention to the categories with higher confidence and reduce the redundant dependencies among the nonexistent classes. To our best knowledge, this is the first application of masked attention in modeling inter-class relationships.
- We design a plug-and-play module, namely the Semantic Disentanglement Module (SDM), to disentangle the high-level semantic feature into a set of category-relevant embeddings for each image by locating the attention region of each category.
- We conduct comprehensive experiments to verify the effectiveness of the proposed approach. On three widely used multi-label aerial scene image recognition benchmarks including UC-Merced Multi-label, AID Multi-label, and MLRSNET, our models consistently have state-of-the-art results.

The rest of this paper is structured as follows. Section 2 gives the related works. Section 3 demonstrates the details of the structure and relative setting of the proposed method. Section 4 is devoted to the discussion of the experiments. Section 5 gives the ablation studies. Section 6 gives the qualitative results. Section 7 analyzes the experimental results and discusses the difference among the proposed method and the previous methods. Section 8 presents the conclusion.

## 2. Related Work

Multi-label aerial scene image classification plays a vital role in the imagery interpretation for remote sensing images. Recently, significant progress has been achieved in multi-label image classification tasks. The proposed approaches for multi-label image classification can be roughly categorized into two aspects, i.e., spatial information and label correlations.

### 2.1. Spatial Information

The utilization of spatial information is key to improving the performance of visual recognition, especially multi-label image recognition. The objects in different locations in an image are usually present at different scales. Pioneering works [9,10] introduce the regional proposal technique from the object detection field and transfer the task into the SLIC tasks in each generated proposal. However, accurate regional proposals need the supervision of extra object-level annotations, which are much more costly than image-level annotations. Hence, some approaches replace the explicit proposals with semantically relevant attentional regions. Wang et al. [11] propose using a spatial Transformer network [31] to ascertain the interest regions corresponding to the semantic labels and predict the scores via Long Short-Term Memory (LSTM) [32]. MCAR [12] presents a two-stream network to recognize multi-category objects from a global image to local regions and designs a multi-class attentional region module to generate a smaller number of attentional regions. Xiong et al. [33] propose a Confounder-Free Fusion Network (CFF-NET) to extract fine-grained deep features from the whole image and provide more multi-grained image information via visual attention. Liang et al. [13] extract multi-scale image features by using multi-scale Graph Convolutional Networks (GCNs).

### 2.2. Label Correlations

Since the co-occurrence of objects in images conforms to the general rules of the real world, mining the label distribution as prior knowledge of subsequent classification enhances the performance of multi-label image classification.

#### 2.2.1. RNN-Based Methods

The CNN-RNN [14] framework learns label correlations with an LSTM layer. Order-free RNN [15] introduces confidence-ranked LSTM and visual attention to increase the flexibility of the label sequence. Orderless RNN [16] proposes an orderless loss, which can dynamically order the labels based on the prediction of the LSTM model to reduce duplicate prediction. Hua et al. [34] exploit the class attention learning layer to generate category-specific features and use a bi-LSTM to model the label dependency in both directions and produce structured multiple object labels. Ji et al. [17] introduce the attention module to separate the high-level semantic features by channel and sequentially predict labels via an LSTM network. Wang et al. [18] propose a Semantic Supplementary Network with Prior Information (SSNP) to first generate prior information by using an LSTM-based prior information network and generate semantic information of the potential labels via a semantic supplementary module.

#### 2.2.2. Graph-Based Methods

Compared to the sequential methods, graph-based approaches have attracted more attention. Early works coped with such a dependency via probabilistic graph models, such as the cyclic directed graphical model [35] and tree-structured graph model [19]. The Graph Convolutional Network (GCN) [22] has been proven to be a more effective tool in modeling label correlations. Chen et al. [36] propose to construct a directed graph based on statistical label co-occurrence and word embedding of labels and propagate the inter-class information by the GCN. Besides the relationships among pairwise labels, Wu et al. [37] model the high-order semantic dependencies via hypergraph neural networks. The construction of the graph usually needs the pre-defined correlation matrix, which is set up with the statistic in the dataset (e.g., statistical label co-occurrence). The static graph cannot accurately represent the characteristic of an individual image. To tackle this problem, Ye et al. [38] use high-level spatial features to build an updatable graph via a Dynamic Graph Convolutional Network (D-GCN) module. In the remote sensing field, graph-based methods have been widely researched. Chaudhuri et al. [39] propose to build an image neighborhood graph via a semi-supervised graph-theoretic method for multi-label remote sensing image retrieval. Tan et al. [40] convert low-rank representation to a feature-based graph and a semantic-

based graph, respectively. Similarly, Zhang et al. [41] build an image feature graph and a semantic graph. The difference is they regularize dual graphs via a non-negative matrix tri-factorization-based collaborative filtering framework. Li et al. [42] construct a scene graph with the extracted feature and mine the spatio-topological relationships of the scene graph via a Graph Neural Network (GNN). Li et al. [43] propose a CM-GM framework to align the image feature and label feature via a GCN, and then, the aligned features are fed into bi-LSTM to predict the image-level labels.

### 2.2.3. Transformer-Based Methods

Transformer [24] is a conspicuous architecture that exploits self-attention to model positionwise relationships among the elements in a long sequence. Unlike CNNs or RNNs, Transformer demonstrates a greater capability for long-range modeling and adaptability to multiple domains, whether NLP or CV. Recently, Transformer has been explored in multi-label classification to model the label correlations [28,44]. Lanchantin et al. [44] propose to capture the dependencies among a set of label embeddings and adaptively combine spatial features via a Transformer encoder. Rather than label embeddings, Chen et al. [28] send category-specific activation maps into the Transformer encoder to exploit the relationships among categories. In the remote sensing field, Deng et al. [45] jointly train a CNN and a vision Transformer to combine the local and global features. Tan et al. [29] ratiocinate the inter-class relation matrix in a Transformer-based SRBM module to generate a robust semantic relationship category representation. Transtl [46] aims to adaptively locate the interested region of each label via one or more STLD modules.

Additionally, there have been other studies of multi-label recognition in the remote sensing field. Wang et al. [47] propose a Multi-Label Semantic Feature Fusion (MLSFF) framework, which consists of a multi-label semantic attribute extractor to extract multi-label semantic attributes and two cross-modal semantic feature fusion operators that fuse the extracted semantic attributes and the image feature extracted by the convolutional neural network. Yu et al. [48] propose a Self-Correction Integrated Domain Adaptation (SCIDA) method for automatic multilabel learning, including a Labelwise Self-Correction (LWC) module to better explore underlying label correlations.

## 3. Methods

In this section, we introduce a novel Multi-Label Image Classification (MLIC) framework named Semantic-driven Masked Attention Transformer (S-MAT), which provides a Transformer-based solution to make use of inter-class relationships to improve classification performance. This section consists of four parts. We first review Transformer in Section 3.1. Then, we introduce the Semantic Disentanglement Module (SDM) in Section 3.2 and the Masked Attention Transformer (MAT) in Section 3.3. In the end, we briefly describe the final classification and loss function in Section 3.4.

### 3.1. Recap of Transformer

The standard Transformer [24] architecture is a typical encoder–decoder architecture. This work is based on the Transformer encoder; thus, we shall introduce the Transformer encoder in the following. The Transformer encoder is a multi-level architecture, in which each layer comprises two key components, namely multi-head self-attention module and feed-forward network module. In the area of natural language processing, the conventional Transformers build relationships among different semantic words in the input language sentences from global perspectives. In other areas, the non-serialized input data need to be preprocessed into a sequence. For example, ViT [6] proposes to cut the image into multiple patches and flatten them into a sequence. Since the sequence loses positional information on the input, position embedding is introduced to preserve the relative position of each element in the sequence.



Given a sequence  $X \in \mathbb{R}^{l_x \times D}$ , where  $l_x$  is the length of  $X$ , they are converted into queries  $Q$ , keys  $K$ , and values  $V$  by the fully connected layers.

$$Q = XW^Q, K = XW^K, V = XW^V, \tag{1}$$

where  $W^Q \in \mathbb{R}^{D \times d^k}$ ,  $W^K \in \mathbb{R}^{D \times d^k}$ , and  $W^V \in \mathbb{R}^{D \times d^v}$  are the learnable parameters for channel transformation.  $d^k$  and  $d^v$  are the channel numbers of the key and value. In this paper, we set  $d^k = d^v = 512$ . The standard dot-product attention with the residual path is defined in Equation (1).

$$A(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V + Q. \tag{2}$$

Then, the updated query embedding  $Q$  is subjected to a fully connected Feed-Forward Network (FFN) to perform a nonlinear mapping. The FFN consists of two linear transformations with a GELU activation in between. The FFN with a residual path is as follows:

$$\text{FFN}(Q) = \text{GELU}(QW_1 + b_1)W_2 + b_2 + Q, \tag{3}$$

where  $W$  and  $b$  stand for the weight matrices and the bias. The subscripts represent different fully connected layers. In this paper, we set the dimension of  $W_1, W_2, b_1$ , and  $b_2$  as  $W_1 \in \mathbb{R}^{512 \times 2048}$ ,  $W_2 \in \mathbb{R}^{2048 \times 512}$ ,  $b_1 \in \mathbb{R}^{2048}$ , and  $b_2 \in \mathbb{R}^{512}$ , respectively.

In our approach, we replace the standard dot-product attention operator with a masked attention operator based on the meta-architecture mentioned above. The overview of our S-MAT framework is presented in Figure 2. It consists of four main parts: (i) high-level feature extraction of the input image via a pre-trained backbone, (ii) construction of label embeddings in the Semantic Disentanglement Module (SDM), (iii) relationship modeling and embedding refinement in a Masked Attention Transformer (MAT), and (iv) computing the final prediction logits for each category. Note that our method can be attached to any backbone without intrusive modifications. The detail will be introduced in the next part.

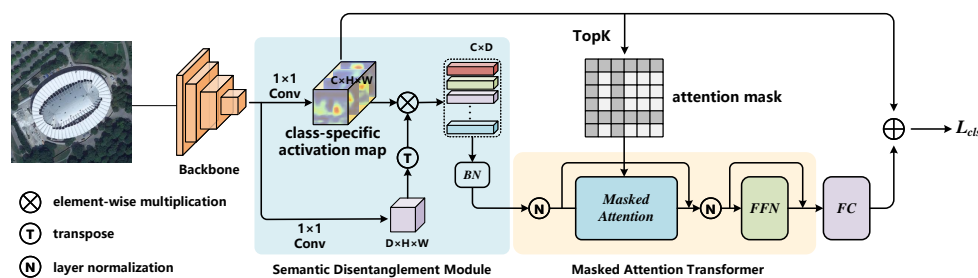


Figure 2. The overall framework of our proposed method.

### 3.2. Semantic Disentanglement Module

Given an image  $I \in \mathbb{R}^{3 \times H_0 \times W_0}$ , where  $H_0$  and  $W_0$  are the height and the width of  $I$ , high-level feature map  $X \in \mathbb{R}^{D' \times H \times W}$  is extracted from the backbone. In this paper, the label embeddings are constructed by disentangling the image feature map  $X$ . The disentanglement is divided into two steps: (1) generation of class-specific activation  $M$ ; (2) matrix product of  $X$  and  $M$ . Class-specific activation represents the probability of a label appearing at each spatial location, and its generation can be formulated as follows:

$$M = \sigma(\varphi_m(X)) \in \mathbb{R}^{C \times H \times W}, \tag{4}$$

where  $\varphi_m(\cdot)$  denotes the  $1 \times 1$  convolution layer to transform the dimension of  $X$  from  $D'$  to the number of classes  $C$  in the current dataset and  $\sigma(\cdot)$  is the *Sigmoid*( $\cdot$ ) function. In this paper,  $D'$  was set to 2048. Each element  $m_{ij}^c$  in class-specific activation  $M \in \mathbb{R}^{C \times H \times W}$  represents the probability of specific category  $c$ 's presence in the feature map  $X$  at  $(i, j)$ . We

adopt Global Max Pooling (GMP) on  $M$  to generate the predictive logit  $y_a = [y_a^1, y_a^2, \dots, y_a^C]$ , which is constrained by the loss demonstrated in Section 3.4 for learning a more accurate representation  $M$ . Hence, the label embeddings are constructed by the product of  $M$  and the transformed feature map as follows:

$$E = \mathcal{R}(M)\varphi_c(\mathcal{R}(X)^\top) \in \mathbb{R}^{C \times D}, \quad (5)$$

where  $\varphi_c$  denotes the  $1 \times 1$  convolution layer to reduce the dimension of  $X$  from  $D' = 2048$  to  $D = 512$  and  $\mathcal{R}$  represents the reshape operation, which squeezes the spatial dimensions  $H$  and  $W$  into one dimension  $HW$ . Intuitively, each category  $c$  selects the interested region in the transformed feature map to combine. As a consequence, each embedding  $e_c$  aggregates the corresponding spatial feature and semantic information.

### 3.3. Masked Attention Transformer

Transformer has proven its outstanding ability to model long-range dependencies. In particular, the built-in mask matrix, which restricts the scope of attention, makes Transformer a perfect choice for modeling label relationships. The mask matrix in Transformer was originally intended to eliminate the effect of padding on the sequence in training or avoid exposing the decoder to predictive content in machine translation. In Mask2Former, the mask matrix is exploited to realize local attention by constraining the attention to the foreground region, instead of the full feature map. As mentioned in Section 1, one crucial problem in multi-label classification is removing the redundant part and obtaining more accurate label dependencies. To solve this problem, we make the first attempt to introduce masked attention into multi-label classification to mask the redundant label dependencies, and the attention is confined to the categories with higher confidence. The proposed masked attention is shown in Figure 3.

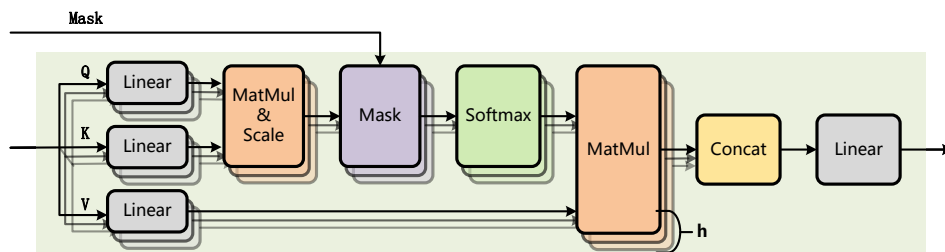


Figure 3. Illustration of the masked attention.

Suppose the ground-truth label set of the input image is  $C_1$ , which contains  $n_1$  labels, and the set of nonexistent labels is  $C_2$ , which contains  $n_2$  labels.  $D(X, Y)$  represents the relationship among the labels in both label sets  $X$  and  $Y$ . We believe that there are higher relations among the labels in  $C_1$ , namely  $D(C_1, C_1) = \{1\}_{n_1 \times n_1}$ . On the contrary,  $D(C_1, C_2) = \{0\}_{n_1 \times n_2}$ . However,  $D(C_2, C_2)$  is uncertain and redundant. If we filter out the redundant part, we can obtain more accurate label relationships. In this paper, we simply judge the confidence of each label according to  $y_a$  and then generate the mask  $\mathcal{M}$ . Since  $y_a$  is not completely accurate, we retain some redundancy in the generation of masks by adjusting the proportion to be filtered out.

To generate the attention mask  $\mathcal{M}$ , we first obtain  $I$ , the index set of the top- $k$  prediction in  $y_a$ , via the *topK* operator in PyTorch. Then, the attention mask  $\mathcal{M}$  is

$$\mathcal{M}(x, y) = \begin{cases} -\infty & \text{if } x, y \notin I \\ 1 & \text{otherwise} \end{cases}. \quad (6)$$

Thus, the masked attention  $A_m$  can be defined as follows:

$$A_m = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}} + \mathcal{M}\right)V + Q. \quad (7)$$

Extending the attention mechanism to a multi-head version enables the mechanism to consider different aspects of the label relationships. The masked multi-head attention ( $MHA_m$ ) mechanism is the cascade of Equation (7), and its definition is shown as follows:

$$Q_i = XW_i^Q, K_i = XW_i^K, V_i = XW_i^V, \quad (8)$$

$$Z_i = A_m(Q_i, K_i, V_i), i = 1, \dots, h, \quad (9)$$

$$MHA_m(Q, K, V) = \text{Concat}(Z_1, Z_2, \dots, Z_h)W^O, \quad (10)$$

The symbols  $W_i^Q \in \mathbb{R}^{D \times d^k}$ ,  $W_i^K \in \mathbb{R}^{D \times d^k}$ , and  $W_i^V \in \mathbb{R}^{D \times d^v}$  are the parameter matrices, where  $d^k = d^v = D/h$ .  $W^O \in \mathbb{R}^{d^v \times D}$  is the output transform matrix.  $h$  is the number of heads, and  $Z_i$  is the output of each attention head.

Our masked Transformer encoder is a multi-layer architecture in which each layer consists of a masked multi-head attention mechanism and a Feed-Forward Network (FFN). With the label embedding from the previous layer  $E_{l-1}$ , each Transformer encoder layer exploits the label relationships and updates the label embedding  $E_{l-1}$  as follows:

$$E_l^{(1)} = E_{l-1} + MHA_m(\hat{E}_{l-1}, \hat{E}_{l-1}, E_{l-1}), \quad (11)$$

$$E_l = E_l^{(1)} + FFN(E_l^{(1)}), \quad (12)$$

The symbol  $\hat{\cdot}$  means the feature modified by adding position encoding.  $E^{(1)}$  is the intermediate variable.

### 3.4. Final Classification and Loss Function

After being refined by the  $l$  Transformer encoder layer, we can obtain the final label embeddings  $E_l \in \mathbb{R}^{C \times D}$  and project them to the logit  $y_r = [y_r^1, y_r^2, \dots, y_r^C]$  via a linear projection layer:

$$y_r = \sum_{c=1}^C (E_l^c W^T) + b, \quad c = 1, 2, \dots, C, \quad (13)$$

where  $W = [W_1, W_2, \dots, W_C] \in \mathbb{R}^{C \times D}$  and  $b \in \mathbb{R}^C$ . Then, the final label confidence  $\hat{y}$  is obtained by elementwise summing up  $y_a$  and  $y_r$ .

$$\hat{y} = \sigma(y_a + y_r), \quad (14)$$

where  $\sigma(\cdot)$  is the *Sigmoid*( $\cdot$ ) function. The ground-truth label of the input image is  $y = [y^1, y^2, \dots, y^C]$ , where  $y^i = 0, 1$  denotes the absence or presence of label  $i$  in the image. The whole framework is trained in an end-to-end manner with the traditional multi-label classification loss as follows:

$$L(y, \hat{y}) = \sum_{c=1}^C y^c \log(\sigma(\hat{y}^c)) + (1 - y^c) \log(1 - \sigma(\hat{y}^c)), \quad (15)$$

where  $\sigma(\cdot)$  is the *Sigmoid*( $\cdot$ ) function.

## 4. Experiment

### 4.1. Dataset

To verify the effectiveness of our proposed method, we compared the proposed method with the state-of-the-art on three popular multi-label remote sensing image datasets: UC-Merced multi-label [39], AID Multi-label [49], and MLRSNET [50].



#### 4.1.1. UC-Merced Multi-Label

The UC-Merced land use dataset is an aerial image dataset, which was originally built in a single-label style. The UC-Merced Multi-label dataset is a duplicate of the UC-Merced dataset and assigns all 2100 images to 17 newly defined object labels. There are 1–7 categories of objects in each image, with a size of  $256 \times 256$ . Following the division in [34,49], 80% of the image samples were exploited in the training phase, and the rest were used in the testing phase.

#### 4.1.2. AID Multi-Label

The AID Multi-label dataset is a subset of the AID dataset. This subset contains 3000 aerial images, which are assigned multiple object labels in 17 categories. AID is a large-scale aerial image dataset that contains 10,000 high-resolution aerial images collected from Google Earth imagery. All the images in AID Multi-label have a size of  $600 \times 600$ . We followed the 80-20 training–test split in [34,51].

#### 4.1.3. MLRSNet

MLRSNet is a multi-label remote sensing dataset, which contains 109,161 high-spatial-resolution optical satellite images captured from different perspectives of the world. The images in MLRSNet are annotated into 46 categories, and the number of sample images in a category varies from 1500 to 3000. The dataset covers 60 predefined categories, with one or more categories per image (up to 13). The resolution of each image ranges from 0.1 m to 10 m, and the size is fixed to  $256 \times 256$ . We followed the standard split used in [50]. A total of 20% of samples were randomly selected for training and 80% for testing.

### 4.2. Evaluation Metrics

In our experiment, we adopted the overall precision (OP), recall (OR) and per-category precision (CP), recall (CR), and F1-measure (CF1) for further comparison. For each image, we assigned each class as positive if its prediction probability was greater than 0.5 and then compared them with the ground-truth labels. The overall precision (OP), recall (OR) and per-category precision (CP), recall (CR), and F1-measure (CF1) are computed as follows:

$$\begin{aligned} \text{OP} &= \frac{\sum_i M_c^i}{\sum_i M_p^i}, & \text{OR} &= \frac{\sum_i M_c^i}{\sum_i M_g^i}, \\ \text{CP} &= \frac{1}{C} \sum_i \frac{M_c^i}{M_p^i}, & \text{CR} &= \frac{1}{C} \sum_i \frac{M_c^i}{M_g^i}, \\ \text{CF1} &= \frac{2 \times \text{CP} \times \text{CR}}{\text{CP} + \text{CR}}, \end{aligned} \quad (16)$$

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.

Note that these results may be affected by the threshold. Among these metrics, the CF1 is more critical and comprehensive, considering both recall and precision.

### 4.3. Implementation Details

In this paper, all the experiments were conducted on a workstation with an AMD EPYC 7543 32-Core Processor, one 32 GB RAM, and a 24 GB RTX A5000 GPU. The base deep learning framework was PyTorch 1.9.0 with Python 3.8 and Cuda 11.1. Following the recent works, we adopted ResNet50 and ResNet101 [4] as our backbone, which were pre-trained on ImageNet [52] for the initialization of the parameters. The input image was first resized to  $224 \times 224$  for the UC-Merced Multi-label dataset and MLRSNet dataset and  $512 \times 512$  for the AID Multi-label dataset. To make a quick convergence, we followed [29], adopting the model trained on Pascal VOC 2012 [53] as the pre-trained model for the UC-Merced

Multi-label dataset and the AID Multi-label dataset. RandAugment [54] and Cutout [55] were adopted for data augmentation. The size of the output feature from the backbone was  $H \times W \times D' = 14 \times 14 \times 2048$ . We reduced the channel of label embeddings  $E$  from  $D' = 2048$  to  $D = 512$ . For each layer of Transformer, we exploited 4 attention heads, and the dimensions of the attention head and feed-forward network were set to  $D/4 = 128$  and  $4D = 2048$ , respectively. We did not use dropout [56] in each layer of Transformer. The batch size of each GPU was 64. Adam [57] was chosen as the optimizer to train the model for 80 epochs, with the weight decay of  $10^{-2}$ ,  $(\beta_1, \beta_2) = (0.9, 0.999)$ , and a learning rate of  $10^{-4}$ .

#### 4.4. Comparison with State-of-the-art Methods

##### 4.4.1. Performance on UC-Merced Multi-Label

We report the results of experiments on the UC-Merced Multi-label dataset. Our proposed method is superior to the previous methods listed in Table 1 in most of the metrics. Generally, CF1 is the primary metric since the others may be greatly affected by the chosen threshold. To be specific, S-MAT-ResNet50 improves CF1 by 9.7% compared to the baseline ResNet50 model. In comparison with MLRSSC-CNN-GNN, which is an effective GNN-based method, our S-MAT-ResNet50 achieves an improvement of 2.82% in CF1, 0.80% in CP, and 0.56% in CR. Both ResNet50-SR-Net and our method are Transformer-based methods. Our proposed MAT-ResNet50 beats ResNet50-SR-Net in the CF1, CP, CR, and OR metrics. The results show the effectiveness of our method.

**Table 1.** Comparisons of our method with previous state-of-the-art methods on the UC-Merced Multi-label dataset. Among all the metrics, CF1 is the primary metric. The bold means the best performance. All metrics are in %.

Method	CF1	CP	CR	OP	OR
ResNet50 [4]	79.51	88.52	78.91	80.70	81.97
ResNet-RBFNN [58]	80.58	86.21	83.72	79.92	84.59
CA-ResNet-BiLSTM [34]	81.47	86.12	84.26	77.94	89.02
CM-GM-N-R-BiLSTM [43]	81.58	88.57	85.20	81.60	89.65
AL-RN-ResNet50 [49]	86.76	<b>88.81</b>	87.07	86.12	84.26
MLRSSC-CNN-GNN [42]	86.39	87.11	88.41	-	-
ResNet50-SR-Net [29]	88.67	87.96	89.40	<b>93.52</b>	91.51
<b>S-MAT-ResNet50 (Ours)</b>	<b>89.21</b>	87.97	<b>89.96</b>	92.94	<b>92.38</b>

##### 4.4.2. Performance on AID Multi-Label

Table 2 shows the quantitative results of the AID Multi-label dataset. The metrics are identical to the ones in the UC-Merced Multi-label dataset. As shown in Table 2, our method achieves the best CF1 score, which is more vital than other indicators. Specifically, our S-MAT-ResNet50 improves the CF1 score of ResNet50, ResNet-RBFNN, CA-ResNet-BiLSTM, AL-RN-ResNet50, MLRSSC-CNN-GNN, and ResNet50-SR-Net by 4.67%, 7.13%, 3.27%, 2.18%, 2.26%, and 0.93%, respectively. Moreover, the performance of our S-MAT-ResNet50 on CP and OP outperforms another Transformer-based approach, ResNet50-SR-Net, by 2.75% and 0.97%, respectively. These results convincingly prove the effectiveness of our proposed method.

##### 4.4.3. Performance on MLRSNet

The comparison with the previous methods on MLRSNet is reported in Table 3. We conducted experiments on two baseline CNN backbones, ResNet50 and ResNet101. As we can observe, the proposed S-MAT-CNN method outperforms the listed competitors in all metrics. It is noteworthy that, compared to the state-of-the-art method CNN-SR-Net, the proposed method improves the CF1 by 1.10% and 1.31% with ResNet50 and ResNet101 as the backbone, respectively.

**Table 2.** Comparisons of our method with previous state-of-the-art methods on the AID Multi-label dataset. Among all the metrics, CF1 is the primary metric. The bold means the best performance. All metrics are in %.

Method	CF1	CP	CR	OP	OR
ResNet50 [4]	86.23	89.31	85.65	72.39	52.82
ResNet-RBFNN [58]	83.77	82.84	88.32	60.85	70.45
CA-ResNet-BiLSTM [34]	87.63	89.03	88.95	79.50	65.60
AL-RN-ResNet50 [49]	88.72	91.00	88.95	80.81	71.12
MLRSSC-CNN-GNN [42]	88.64	89.83	90.20	-	-
ResNet50-SR-Net [29]	89.97	89.42	<b>90.52</b>	87.24	<b>82.25</b>
<b>S-MAT-ResNet50 (Ours)</b>	<b>90.90</b>	<b>92.17</b>	89.69	<b>88.21</b>	80.70

**Table 3.** Comparisons of our method with previous state-of-the-art methods on the MLRSNet dataset. \* denotes our implementation. Among all the metrics, CF1 is the primary metric. The bold means the best performance. All metrics are in %.

Method	CF1	CP	CR	OP	OR
ResNet50 [4]	75.30	-	-	-	-
ResNet50 * [4]	81.35	80.85	81.56	82.19	82.70
ResNet50-SR-Net [29]	87.21	87.08	87.34	88.79	86.73
<b>S-MAT-ResNet50 (Ours)</b>	<b>88.31</b>	<b>87.80</b>	<b>88.79</b>	<b>90.93</b>	<b>91.02</b>
ResNet101 [4]	76.18	-	-	-	-
ResNet101 * [4]	81.89	81.42	82.03	82.65	82.89
ResNet101-SR-Net [29]	87.55	87.84	87.26	89.41	87.48
<b>S-MAT-ResNet101 (Ours)</b>	<b>88.86</b>	<b>88.67</b>	<b>88.93</b>	<b>91.21</b>	<b>91.44</b>

## 5. Ablation Studies

In this section, we perform exhaustive experiments to analyze the essential components of our proposed method. Firstly, we analyzed the contribution of each component in our S-MAT framework. Secondly, the effects of different settings of the Transformer encoder were analyzed, including the number of attention heads, the number of encoder layers, and different position encoding. Thirdly, we discuss the settings of the generation and application of the mask in masked attention, including the selection of  $k$  and the position to apply masked attention. For simplicity, in this section, we shall abbreviate “UC-Merced Multi-label” and “AID Multi-label” to “UC-Merced” and “AID” in all the tables, respectively.

### 5.1. Contributions of the Proposed Method

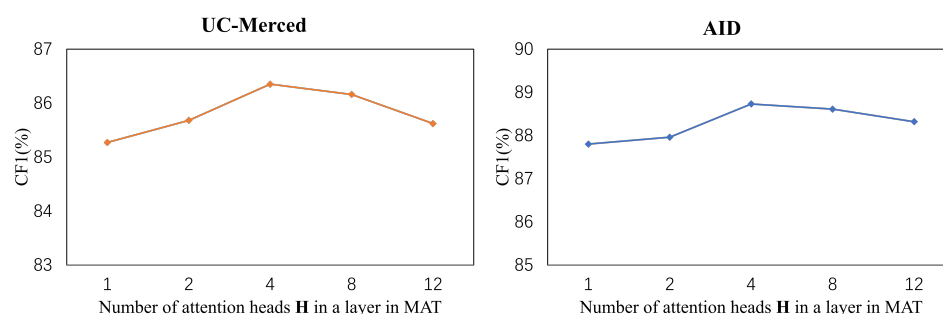
To investigate the effectiveness of each component in our proposed method, we gradually removed SDM and MAT from the complete framework. Meanwhile, to comprehensively explore the contribution of the SDM, we studied the effect of  $y_a$ . We first removed MAT to evaluate the SDM by feeding the raw label embeddings output from the SDM into the final linear projection layer. Then, we replaced the SDM with an MLP layer to evaluate the contribution of the SDM. The result is shown in Table 4. On three popular benchmark datasets, the SDM and MAT significantly promote the baseline, respectively, which convincingly proves the effectiveness of the SDM and MAT. By observing the first three rows in the table, it can be found that the promotions of the SDM are from two aspects. On the one hand, the high-level features from the backbone are disentangled into more discriminative label embedding features. On the other, when the prediction  $y_a$  takes effect, the constraints from the loss interpreted in Equation (15) enable the SDM to generate more accurate class-specific activation map. Moreover, we found that the combination of the SDM and MAT maximizes the performance of CF1 (+9.70% on UC-Merced, +4.67% on AID, +6.96% on MLRSNet), which demonstrates that the SDM and MAT are complementary.

**Table 4.** Ablation studies on the essential components of our proposed method. The symbol  $\checkmark$  represents the component in this column is in use. The red font denotes the improvement over the baseline. The bold means the best performance.

Component		Prediction	CF1		
SDM	MAT	$y_a$	UC-Merced	AID	MLRSNet
-	-	-	79.51	86.23	81.35
$\checkmark$	-	-	82.35 $\uparrow$ 2.84	87.16 $\uparrow$ 0.93	83.46 $\uparrow$ 2.11
$\checkmark$	-	$\checkmark$	84.18 $\uparrow$ 4.67	87.78 $\uparrow$ 1.55	84.71 $\uparrow$ 3.36
-	$\checkmark$	-	85.61 $\uparrow$ 6.10	88.26 $\uparrow$ 2.03	85.93 $\uparrow$ 4.58
$\checkmark$	$\checkmark$	-	87.59 $\uparrow$ 8.08	90.14 $\uparrow$ 3.91	86.48 $\uparrow$ 5.13
$\checkmark$	$\checkmark$	$\checkmark$	<b>89.21</b> $\uparrow$ 9.70	<b>90.90</b> $\uparrow$ 4.67	<b>88.31</b> $\uparrow$ 6.96

### 5.2. Number of Attention Heads

We demonstrated the effectiveness of each component in our proposed approach. In this part, we explore the performance curve of our proposed method by varying the number of attention heads. To avoid the effect of masked attention, we exploited standard dot-product attention instead of masked attention in this part. We set the number of layers in Transformer  $L = 1$  and gradually increased the number of attention heads  $H$  from 1 to 12. The results are shown in Figure 4. The performance curves across UC-Merced Multi-label and AID Multi-label are similar. Appropriately increasing the  $H$  of multi-head attention enhances the performance of CF1. However, when there are too many attention heads, the performance drops since not enough features are assigned to each attention head. When the number of attention heads  $H = 4$ , the performance of CF1 reaches its summit.



**Figure 4.** The results of the ablation studies on the number of attention heads in a layer in MAT.

### 5.3. Number of Layers in Transformer

To evaluate the effect of the different number of layers  $L$  in Transformer on CF1, we changed  $L$  from 1 to 6. To avoid the effect of masked attention, we exploited standard dot-product attention instead of masked attention in this part. Following the best setting in the last part, the number of attention head  $H$  was set to 4. In Figure 5, we show the result of varying the number of layers in Transformer. The results show that the performance of CF1 ascends to the peak when three layers are stacked. By stacking more layers, the performance drops continually on both the UC-Merced and AID datasets.

### 5.4. Position to Apply Masked Attention

In this part, we attempt to figure out the best position to apply our masked attention by replacing the standard attention with our masked attention in different layers. As shown in Table 5, we applied masked attention in (1) the first layer, (2) the second layer, (3) the last layer, and (4) all layers, respectively. The results show that, compared to standard dot-product attention, the application of masked attention in any layer improves the performance of CF1. Moreover, when the masked attention is applied in all layers, we can obtain the best performance. Compared to the model without masked attention

(Line 0), our proposed masked attention provides +1.38%, +0.95%, and +1.92% CF1 on UC-Merced, AID, and MLRSNet, respectively, without introducing any extra parameters.

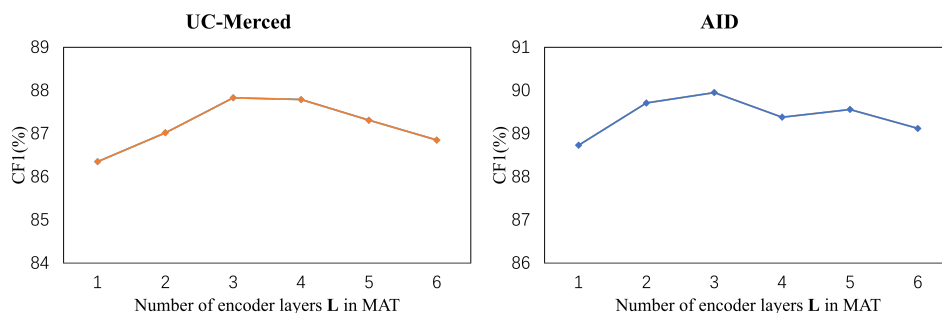


Figure 5. The results of ablation studies on the number of encoder layers in MAT.

5.5. Selection of k in the Generation of the Mask

The selection of k is a fundamental setting in the generation of the mask. With an increasing k, more label dependencies computed in the self-attention operator are filtered out. The results are shown in Figure 6. Note that, if k = 0, no label dependency is available and the model will not converge. Thus, we do not show any result for k = 0 in Figure 6. In this part, the number of layers in MAT L was set to 1. When k was reducing from 100 to 25, our model obtained an increasing CF1. However, when k was less than 25, the CF1 dropped since some categories with high confidence would be ignored as well.

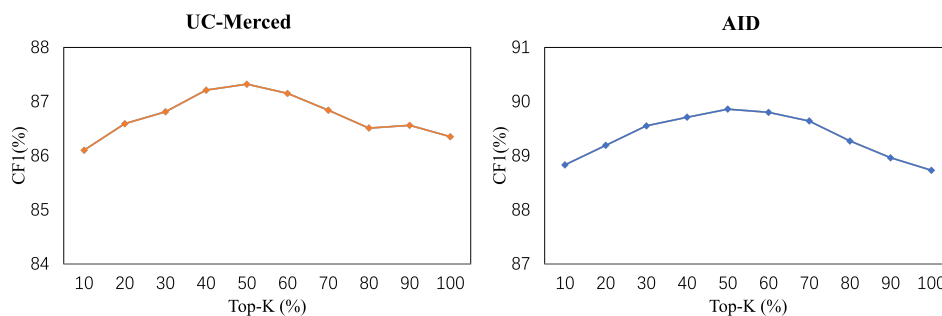


Figure 6. The results of ablation studies on the selection of k in the generation of the mask in MAT.

Table 5. Ablation studies on the position to apply masked attention. The symbol ✓ represents that the masked attention is in use in this layer. The bold means the best performance.

Masked Attention			CF1		
Layer 1	Layer 2	Layer 3	UC-Merced	AID	MLRSNet
-	-	-	87.83	89.95	86.39
✓	-	-	88.57	90.26	87.42
-	✓	-	88.12	90.08	86.81
-	-	✓	88.85	90.43	87.70
✓	✓	✓	<b>89.21</b>	<b>90.90</b>	<b>88.31</b>

5.6. Position Embedding

Since position embedding retains the relative position of the input sequence in the original Transformer, we attempted to figure out its effectiveness in our method. As shown in Table 6, when the label embeddings are elementwise summed up with the position embedding, the performance of CF1 improves around 0.1% on MLRSNet. Meanwhile, position embedding barely contributes to the performance of UC-Merced and AID. We speculate that since there are more categories and more complicated label relationships in MLRSNet, the model with position embedding is slightly better than the one without it.



**Table 6.** Ablation studies on position embedding. The symbol  $\checkmark$  represents that position embedding is in use. The bold means the best performance.

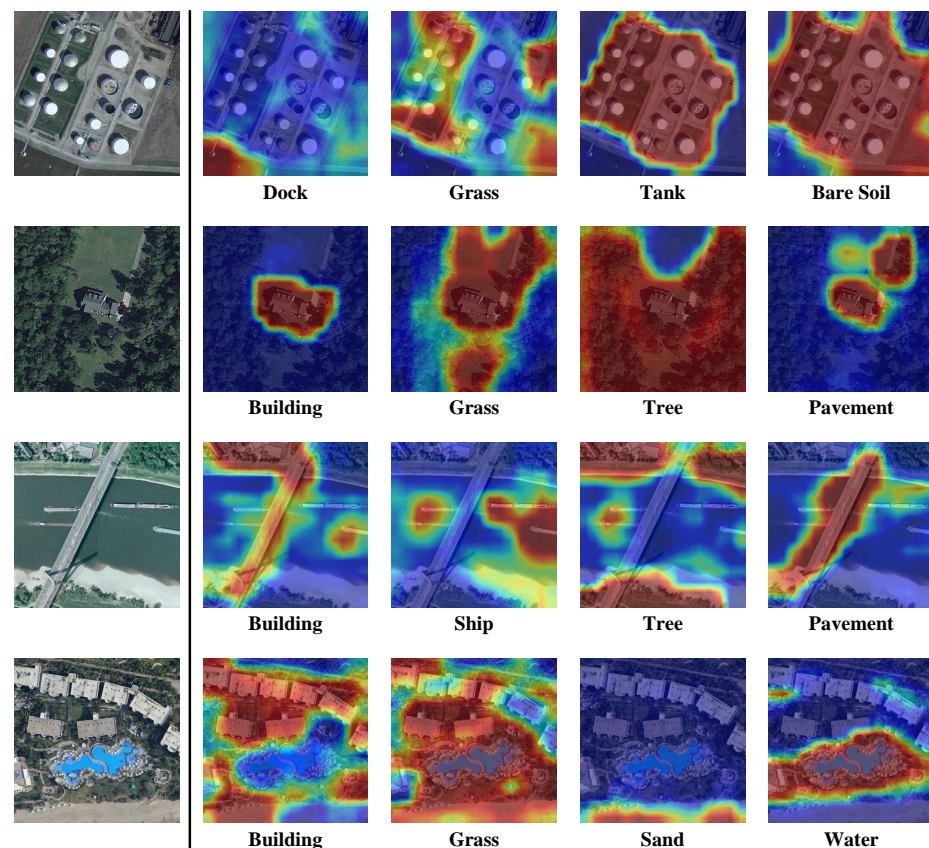
Method	CF1		
Position Embedding	UC-Merced	AID	MLRSNet
-	89.19	90.89	88.22
$\checkmark$	<b>89.21</b>	<b>90.90</b>	<b>88.31</b>

## 6. Visualization

In this section, qualitative results are presented to further reveal the effectiveness of the proposed models. Specifically, we firstly visualize the class-specific activation map in the SDM to show if the SDM could provide precise localization. Then, we visualize the learned inter-class relationship matrix in MAT to demonstrate the ability to capture the co-existence of objects.

### 6.1. Visualization of Class-Specific Activation Map

To investigate the capability of locating the position of each category in the image, we visualize the class-specific activation map  $M$  in the SDM. Some qualitative results are shown in Figure 7. Each row presents the raw image and the corresponding class-specific activation map of each ground-truth label. We can find that the SDM is able to locate the corresponding region of each category in the image. For instance, the ground-truth labels of the raw image in the third row are “building”, “ship”, “tree”, and “pavement”. Our SDM module can provide accurate localization of these four categories. In addition, the accurate activation map leads to more precise label embeddings for modeling the inter-class dependencies in MAT.



**Figure 7.** Visualization of the class-specific activation map in the SDM on the AID Multi-label dataset.



### 6.2. Visualization of Relationship Matrix in MAT

To further demonstrate the effectiveness of our method, we visualize the relationship matrix in the last layer of MAT. Partial results of the visualization are shown in Figure 8. Each row in Figure 8 consists of a raw image and the visualization of the corresponding relation matrix in MAT. We can observe that our proposed MAT is able to explore and capture the accurate label dependencies. Specifically, for the input image in the first row, the ground-truth labels are (“airplane”, “bare soil”, “buildings”, “cars”, “grass”, “pavement”, “trees”). In the relation matrix on the right, we can find that  $A(\text{trees, airplane})$ ,  $A(\text{pavement, airplane})$ ,  $A(\text{grass, airplane})$ ,  $A(\text{buildings, airplane})$ , and  $A(\text{bare soil, airplane})$  are ranked top in the column of “airplane”. This indicates that the correlations between the label pair (“trees”, “airplane”), (“pavement”, “airplane”), (“grass”, “airplane”), (“buildings”, “airplane”), and (“bare soil”, “airplane”) are higher than the others (greater than 0.7). Meanwhile, for the input image in the second row, the ground-truth labels are (“bare soil”, “buildings”, “cars”, “courts”, “grass”, “pavement”, “trees”). In the relation matrix on the right, we can find that  $A(\text{trees, courts})$ ,  $A(\text{pavement, courts})$ ,  $A(\text{grass, courts})$ ,  $A(\text{cars, courts})$ ,  $A(\text{buildings, courts})$ , and  $A(\text{bare soil, courts})$  are ranked top in the column of “courts”. This indicates that the correlations between the label pair (“trees”, “courts”), (“pavement”, “courts”), (“grass”, “courts”), (“cars”, “courts”), (“buildings”, “courts”), and (“bare soil”, “courts”) are higher than the others (greater than 0.7). These convincingly demonstrate that our proposed method can explore, capture, and exploit the label relationships in specific images.

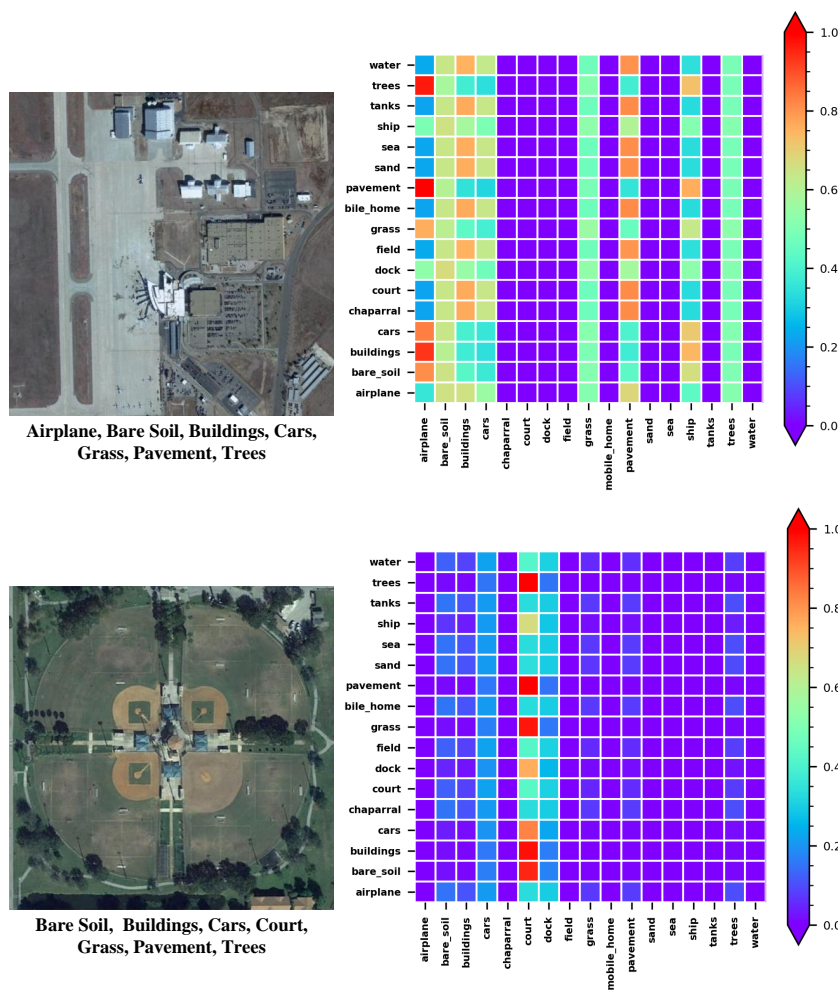


Figure 8. Visualization of the relation matrix in MAT on the AID Multi-label dataset.

## 7. Discussion

In this section, we demonstrate the analyses of the experimental results and differences between our S-MAT and previous methods. As shown in Tables 1 and 2, we can find that S-MAT not only improves the performance of precision and recall, but also keeps them in balance. Consequently, S-MAT always wins the comparison of CF1, which considers both precision and recall. Secondly, the RNN-based methods, such as CA-ResNet-BiLSTM and CM-GM-N-R-BiLSTM, barely surpass the baseline ResNet50. Limited by modeling as the pre-set sequence, RNN-based methods are incapable of capturing the accurate inter-class relationship. Different from the RNN, Transformer can compute the attention matrix among the labels. Therefore, Transformer-based methods, such as ResNet50-SR-Net and our S-MAT-ResNet50, are the superior approaches to model the relationships between pairwise labels. Moreover, compared with SR-NET, in the S-MAT framework, MAT replaces standard dot-product attention with masked attention. The latter can filter out redundancy to obtain a more accurate inter-class relationship. Meanwhile, the SDM disentangles the global feature map to generate the label embeddings and, thus, introduce valuable spatial information for feature representation. As demonstrated in Table 4, the combination of these two modules maximizes recognition performance on the three benchmark datasets.

Figure 9 shows a qualitative example on the AID Multi-label dataset. The first column is the input image. The second column and the third one are the methods and the output predictions, respectively. Green labels denote true positive; red labels denote false positive; gray ones denote false negative. We can find that the baseline ResNet-50 failed to distinguish the objects with similar appearances, such as “trees”, “grass”, and “field” since the model only utilizes the global spatial features. The other methods obtain a lower false positive rate than the baseline by dealing with label dependencies. Those inconspicuous objects, such as “cars” and “dock”, can not be recognized by ResNet-50 and CA-ResNet-BiLSTM. While SR-Net misses the prediction of “dock”, S-MAT can recognize both “cars” and “dock” accurately and obtain a better performance.

It is worth noting that our model leans upon massive annotated large-scale datasets to learn the semantic context and the label dependencies of a visual scene. However, massive annotated data are costly and rare, while most aerial scene images are unlabeled. Therefore, learning the visual feature via self-supervised learning is our research topic in the future.

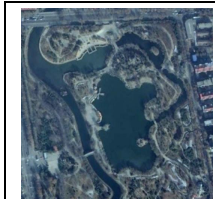
	ResNet-50	bare soil, buildings, cars, dock, grass, pavement, trees, water, field
	CA-ResNet-BiLSTM	bare soil, buildings, cars, dock, grass, pavement, trees, water
	ResNet50-SR-Net	bare soil, buildings, cars, dock, grass, pavement, trees, water
	Ours	bare soil, buildings, cars, dock, grass, pavement, trees, water
	Groudtruth	bare soil, buildings, cars, dock, grass, pavement, trees, water

Figure 9. Qualitative results on the AID Multi-label dataset.

## 8. Conclusions

Multi-label aerial scene image classification is a fundamental task in the computer vision area. Label dependency is a vital factor for multi-label learning. In this paper, we aimed to model a more accurate inter-class relationship by removing the redundant label dependencies among the low confidence categories. We presented a Transformer-based framework S-MAT for multi-label aerial scene image classification. Specifically, we proposed two plug-and-play modules, the Semantic Disentanglement Module (SDM) and Masked-Attention Transformer (MAT). The SDM aims to locate the semantic region of each category and conduct semantic disentanglement to generate label embeddings for MAT. MAT not only captures and models the complicated inter-class relationships in a specific image, but filters out the redundant label dependencies by replacing the standard dot-product attention in the standard Transformer architecture with the proposed masked attention. Especially, the masked attention significantly improves the performance without

introducing additional parameters. Therefore, our S-MAT consistently outperforms the prior works on three widely used and challenging remote sensing image datasets including UC-Merced Multi-label, AID Multi-label, and MLRSNet. In addition, quantitative and qualitative ablation studies and visualizations convincingly proved the effectiveness of the essential components of our method under different factors. In the future, we will focus on self-supervised learning for multi-label aerial scene image classification.

**Author Contributions:** Conceptualization, H.W.; methodology, H.W.; software, H.W.; validation, H.W.; investigation, H.W. and C.X.; resources, H.L.; data curation, C.X.; writing—original draft preparation, H.W.; writing—review and editing, H.W.; visualization, H.W.; supervision, H.L.; project administration, H.L.; funding acquisition, C.X. and H.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Natural Science Foundation of China (Grant No. 62171042, 62102033, 61871039, 62006020, 61906017), the R&D Program of Beijing Municipal Education Commission(KZ202211417048), the Beijing Municipal Commission of Education Project (No.KM202111417001, KM201911417001), the Collaborative Innovation Center of Chaoyang(Grant No. CYXC2203), the Academic Research Projects of Beijing Union University(No.BPHR2020DZ02, ZB10202003, ZK40202101, ZK120202104).

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

S-MAT	Semantic-driven Masked-Attention Transformer
MAT	Masked Attention Transformer
SDM	Semantic Disentanglement Module
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GCN	Graph Convolutional Network
FC	Fully Connected

## References

1. Wan, L.; Liu, N.; Guo, Y.; Huo, H.; Fang, T. Local feature representation based on linear filtering with feature pooling and divisive normalization for remote sensing image classification. *J. Appl. Remote Sens.* **2017**, *11*, 016017.
2. Xu, K.; Huang, H.; Deng, P. Remote sensing image scene classification based on global-local dual-branch structure model. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5.
3. Wang, H.; Gao, K.; Min, L.; Mao, Y.; Zhang, X.; Wang, J.; Hu, Z.; Liu, Y. Triplet-Metric-Guided Multi-Scale Attention for Remote Sensing Image Scene Classification with a Convolutional Neural Network. *Remote Sens.* **2022**, *14*, 2794.
4. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2016; pp. 770–778.
5. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
6. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021; pp. 1–21.
7. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical vision Transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
8. Mahdy, A. A numerical method for solving the nonlinear equations of Emden-Fowler models. *J. Ocean. Eng. Sci.* **2022**. <https://doi.org/10.1016/j.joes.2022.04.019>.
9. Wei, Y.; Xia, W.; Lin, M.; Huang, J.; Ni, B.; Dong, J.; Zhao, Y.; Yan, S. HCP: A flexible CNN framework for multi-label image classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 1901–1907.

10. Yang, H.; Zhou, J.T.; Zhang, Y.; Gao, B.B.; Wu, J.; Cai, J. Exploit bounding box annotations for multi-label object recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2016; pp. 280–288.
11. Wang, Z.; Chen, T.; Li, G.; Xu, R.; Lin, L. Multi-label image recognition by recurrently discovering attentional regions. In Proceedings of the IEEE International Conference on Computer Vision, 2017, Venice, Italy, 22–29 October 2017; pp. 464–472.
12. Gao, B.B.; Zhou, H.Y. Learning to discover multi-class attentional regions for multi-label image recognition. *IEEE Trans. Image Process.* **2021**, *30*, 5920–5932.
13. Liang, J.; Xu, F.; Yu, S. A multi-scale semantic attention representation for multi-label image recognition with graph networks. *Neurocomputing* **2022**, *491*, 14–23.
14. Wang, J.; Yang, Y.; Mao, J.; Huang, Z.; Huang, C.; Xu, W. Cnn-rnn: A unified framework for multi-label image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, 21–26 July 2016; pp. 2285–2294.
15. Chen, S.F.; Chen, Y.C.; Yeh, C.K.; Wang, Y.C. Order-free rnn with visual attention for multi-label classification. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
16. Yazici, V.O.; Gonzalez-Garcia, A.; Ramisa, A.; Twardowski, B.; van de Weijer, J. Orderless recurrent models for multi-label classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13440–13449.
17. Ji, J.; Jing, W.; Chen, G.; Lin, J.; Song, H. Multi-label remote sensing image classification with latent semantic dependencies. *Remote Sens.* **2020**, *12*, 1110.
18. Wang, Z.; Fang, Z.; Li, D.; Yang, H.; Du, W. Semantic supplementary network with prior information for multi-label image classification. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 1848–1859.
19. Li, X.; Zhao, F.; Guo, Y. Multi-label Image Classification with A Probabilistic Label Enhancement Model. In Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence, Quebec City, Quebec, Canada, 2014; pp. 430–439.
20. Li, Q.; Qiao, M.; Bian, W.; Tao, D. Conditional graphical lasso for multi-label image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, 21–26 July 2016; pp. 2977–2986.
21. Chow, C.; Liu, C. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inf. Theory* **1968**, *14*, 462–467.
22. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. In Proceedings of International Conference on Learning Representations, Toulon, France, 24–26 April 2017; pp. 1–14.
23. Chen, Z.M.; Wei, X.S.; Jin, X.; Guo, Y. Multi-label image recognition with joint class-aware map disentangling and label correlation embedding. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; pp. 622–627.
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS), Long Beach, California, USA, December 4–9 2017; pp. 6000–6010.
25. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with Transformers. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
26. Dai, Z.; Cai, B.; Lin, Y.; Chen, J. Up-detr: Unsupervised pre-training for object detection with Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1601–1610.
27. Lin, A.; Chen, B.; Xu, J.; Zhang, Z.; Lu, G.; Zhang, D. DS-TransUNet: Dual Swin Transformer U-Net for Medical Image Segmentation. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–15. <https://doi.org/10.1109/TIM.2022.3178991>.
28. Chen, Z.M.; Cui, Q.; Zhao, B.; Song, R.; Zhang, X.; Yoshie, O. SST: Spatial and Semantic Transformers for Multi-Label Image Recognition. *IEEE Trans. Image Process.* **2022**, *31*, 2570–2583.
29. Tan, X.; Xiao, Z.; Zhu, J.; Wan, Q.; Wang, K.; Li, D. Transformer-Driven Semantic Relation Inference for Multilabel Classification of High-Resolution Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 1884–1901.
30. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention mask Transformer for universal image segmentation. *arXiv* **2021**, arXiv:2112.01527.
31. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial Transformer networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS), Montreal, Canada, December 7–12 2015; Volume 2, pp. 2017–2025.
32. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780.
33. Xiong, W.; Xiong, Z.; Cui, Y. A Confounder-free Fusion Network for Aerial Image Scene Feature Representation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, 1–15. <https://doi.org/10.1109/JSTARS.2022.3189052>.
34. Hua, Y.; Mou, L.; Zhu, X.X. Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification. *ISPRS J. Photogramm. Remote Sens.* **2019**, *149*, 188–199.
35. Guo, Y.; Gu, S. Multi-label classification using conditional dependency networks. In Proceedings of the International Joint Conference on Artificial Intelligence, Barcelona, Spain, 16–22 July 2011; pp. 1300–1305.
36. Chen, Z.M.; Wei, X.S.; Wang, P.; Guo, Y. Multi-label image recognition with graph convolutional networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5177–5186.

37. Wu, X.; Chen, Q.; Li, W.; Xiao, Y.; Hu, B. AdaHGNN: Adaptive Hypergraph Neural Networks for Multi-Label Image Classification. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 284–293.
38. Ye, J.; He, J.; Peng, X.; Wu, W.; Qiao, Y. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020, pp. 649–665.
39. Chaudhuri, B.; Demir, B.; Chaudhuri, S.; Bruzzone, L. Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method. *IEEE Trans. Geosci. Remote Sens.* **2017**, *56*, 1144–1158.
40. Tan, Q.; Liu, Y.; Chen, X.; Yu, G. Multi-label classification based on low rank representation for image annotation. *Remote Sens.* **2017**, *9*, 109.
41. Zhang, J.; Zhang, J.; Dai, T.; He, Z. Exploring weighted dual graph regularized non-negative matrix tri-factorization based collaborative filtering framework for multi-label annotation of remote sensing images. *Remote Sens.* **2019**, *11*, 922.
42. Li, Y.; Chen, R.; Zhang, Y.; Zhang, M.; Chen, L. Multi-label remote sensing image scene classification by combining a convolutional neural network and a graph neural network. *Remote Sens.* **2020**, *12*, 4003.
43. Li, P.; Chen, P.; Zhang, D. Cross-Modal Feature Representation Learning and Label Graph Mining in a Residual Multi-Attentional CNN-LSTM Network for Multi-Label Aerial Scene Classification. *Remote Sens.* **2022**, *14*, 2424.
44. Lanchantin, J.; Wang, T.; Ordonez, V.; Qi, Y. General multi-label image classification with Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16478–16488.
45. Deng, P.; Xu, K.; Huang, H. When CNNs meet vision Transformer: A joint framework for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5.
46. Wu, H.; Li, M.; Liu, Y.; Liu, H.; Xu, C.; Li, X. Transtl: Spatial-Temporal Localization Transformer for Multi-Label Video Classification. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 1965–1969.
47. Wang, S.; Ye, X.; Gu, Y.; Wang, J.; Meng, Y.; Tian, J.; Hou, B.; Jiao, L. Multi-label semantic feature fusion for remote sensing image captioning. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 1–18.
48. Yu, T.; Lin, J.; Mou, L.; Hua, Y.; Zhu, X.; Wang, Z.J. SCIDA: Self-Correction Integrated Domain Adaptation from Single-to Multi-label Aerial Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–13. <https://doi.org/10.1109/TGRS.2022.3170357>.
49. Hua, Y.; Mou, L.; Zhu, X.X. Relation network for multilabel aerial image classification. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 4558–4572.
50. Qi, X.; Zhu, P.; Wang, Y.; Zhang, L.; Peng, J.; Wu, M.; Chen, J.; Zhao, X.; Zang, N.; Mathiopoulos, P.T. MLRSNet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding. *ISPRS J. Photogramm. Remote Sens.* **2020**, *169*, 337–350.
51. Mou, L.; Hua, Y.; Zhu, X.X. Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7557–7569.
52. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
53. Everingham, M.; Eslami, S.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136.
54. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q.V. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 702–703.
55. DeVries, T.; Taylor, G.W. Improved regularization of convolutional neural networks with cutout. *arXiv* **2017**, arXiv:1708.04552.
56. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
57. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015; pp. 1–15.
58. Zeggada, A.; Melgani, F.; Bazi, Y. A deep learning approach to UAV image multilabeling. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 694–698.