Semantic interleaving global channel attention for multilabel remote sensing image classification

Yongkun Liu\textsuperscript{a}, Kesong Ni\textsuperscript{c}, Yuhan Zhang\textsuperscript{d}, Lijian Zhou\textsuperscript{b}\textsuperscript{,} and Kun Zhao\textsuperscript{b}\textsuperscript{,}

\textsuperscript{a}School of Software Engineering, Sun Yat-Sen University, Zhuhai, China; \textsuperscript{b}School of Information and Control Engineering, Qingdao University of Technology, Qingdao, China; \textsuperscript{c}School of Software, Dalian University of Technology, Dalian, China; \textsuperscript{d}School of Computer Science and Technology, Xidian University, Xi’an, China

ABSTRACT
Multilabel remote sensing image classification (MLRSIC) has received increasing research interest. Taking the co-occurrence relationship of multiple labels as additional information helps to improve the overall performance. However, current methods only focus on using it to constrain the final feature which is output from a convolutional neural network (CNN). On the one hand, these methods need to exploit the potential of label correlation in feature representation fully. On the other hand, they increase the label noise sensitivity of the system, resulting in poor robustness. In this paper, a novel method called ‘Semantic Interleaving Global chaNnel Attention’ (SIGNA) is proposed for MLRSIC. First, the label co-occurrence graph is obtained according to the statistical information of the training set and fed into a graph neural network (GNN) to generate optimal semantic feature representations of each label. Next, the semantic features are interleaved with visual features which are extracted by CNNs to guide the overall features of the input image transform from the original feature space to the semantic feature space with embedded label relations. Then, global attention triggered by semantic interleaving is used to emphasize visual features in important channels. Finally, to make SIGNA easier to use and more optimized, multithread SIGNA-based feature adaptive weighting networks are proposed as plug-in blocks to plug into any layers of a CNN. For remote sensing images, better classification performance can be achieved by inserting the plug-in blocks into the shallow layers of CNNs. We conducted extensive experimental comparisons on three data sets: UCM, AID and DFC15. Experimental results demonstrate that the proposed SIGNA achieves superior classification performance compared to state-of-the-art (SOTA) methods. Notes that the codes of this paper will be open to the community for reproducibility research.

1. Introduction
With the development of satellite technology, remote sensing image scene classification has become a very active field. So far, remote sensing images are now widely utilized in various applications, such as urban cartography (Gao et al. 2017; Hong et al. 2019, 2019; Hua, Mou, and Xiang Zhu 2020; Rasti et al. 2020; Zhu et al. 2019), land use determination (Hua, Mou, and Xiang...
Remote sensing image classification plays a key role in the above tasks and has become one of the fundamental tasks in the remote sensing field. Single label remote sensing image classification aims to predict a semantic classification based on the overall information in the remote sensing image (Lin, Belongie, and Hays 2013; Movshovitz-Attias et al. 2015). However, due to the complexity and richness of information in remote sensing images, it is difficult to describe an image with a single label at the macro level. Unlike the single label used in traditional remote sensing classification, MLRSIC uses a series of object labels to describe remote sensing images. This is more meaningful for semantic description and visual understanding of remote sensing images.

Applying label correlation to multilabel image classification is critical for improving MLRSIC performance. Recently, some related methods have been proposed. Hua et al. proposed a method to extract image features through the CNN backbone network, which connects and sends the image features to bidirectional LSTM for classification, and implicitly learns label co-occurrence information (Hua, Mou, and Xiang Zhu 2019b). Chen et al. proposed the ML-GCN method (Chen et al. 2019). The GNN is used to process the input image by modelling the label correlation as a graph. Koda et al. proposed a multilabel classification method based on a structured support vector machine (SVM) using the relationship between labels (Koda et al. 2018). All of the above methods employ CNN to extract visual features, then use various methods to fuse label features and visual features to apply label correlation to image classification. This feature fusion method has some drawbacks.

As shown in (a) of Figure 3 in section 3.1, in the ML-GCN method, the vector of the output of the CNN and several classifier vectors are mapped to one output. Above all, semantic features are explicitly used to guide image feature classification, which results in the networks being highly sensitive to the numerical value and noise of label features. Furthermore, from the perspective of backpropagation, the image features and semantic features exchange gradients only once in the last layer, which indicates that the semantic features are not fully utilized. Lastly, these methods do not consider the properties of remote sensing images.

As shown in Figure 1, there are two significant distinctions between images in remote sensing classification tasks and images in general visual classification tasks. Firstly, in (a), the image pixels correspond to people with yellow skin, blue-white-black clothes, black pants, and white-black helmets, which are very complex. However, in (c), the pixels corresponding to the tree, grass and water have stronger regional consistency in texture.

Figure 1. (a), (b) are the sample images in general classification tasks and their corresponding semantic segmentation maps. (c), (d) are sample images in remote sensing and their corresponding semantic segmentation maps.
and colour. Secondly, the semantic labels of generic images only correspond to part of the pixels, and there is a lot of meaningless background information. However, the remote sensing labels of images correspond to each pixel.

Based on the issues above, this paper extensively explores the correlation between semantic label co-occurrence relationships and visual features in remote sensing images. The main contributions of this paper are as follows:

1. Semantic interleaving encoding is proposed, inspired by interleaved codes in error correction codes (Vucetic and Yuan 2012). This is to solve the problem of poor network robustness caused by the explicit use of semantic features. It can reduce the numerical and noise sensitivity of fixed-label co-occurrence relationships, thereby improving the robustness of the entire network.
2. SIGNA is proposed for the problem that semantic features are not fully utilized. Global attention to image channels is triggered in a new semantic feature space, and the more significant visual features are extracted by using relationships between labels and channels.
3. A plug-in network based on SIGNA is designed to better use the characteristics of remote sensing image data sets. The network is used in the shallow layer of CNN. On the one hand, it can better use the texture and edge features of remote sensing images. On the other hand, it does not ignore any meaningful pixel in the shallow layer.

The rest of the paper is organized as follows. In Section 2, we review the multilabel image classification algorithms in remote sensing, the current research progress of attention mechanism in multilabel classification, and introduce the application of GNN in the direction of image recognition. Section 3 introduces the proposed SIGNA and describes in detail its three modules. Section 4 reports the experimental setup. Section 5 reports the experimental results and analysis. Finally, the conclusion is given in Section 6.

2. Related work

The MLRSIC task is a hot research topic in remote sensing, which can provide a more fine-grained understanding of images. We review related work from the following aspects: traditional methods for MLRSIC, using label relationships and attention mechanisms, and GNN applications in MLRSIC.

2.1. Multilabel classification in remote sensing

In the early multilabel classification research, hand-crafted features (such as colour, texture and visual word bag) describe the image scene and combine it with traditional machine learning methods (Dai et al. 2018; Geng et al. 2015; Koda et al. 2018; Zeggada et al. 2018), such as random forest and support vector machine (SVM). However, the generalization ability of the above research is limited, and it is not easy to express the overall and high-level semantic information. Therefore, the classification performance of these algorithms is limited.

In the field of computer vision, with the establishment of large-scale natural image data sets and the rapid development of deep convolution networks, such as VGG16 (Simonyan,
Using multilabel relationships and attention mechanisms

Introducing label correlation and attention mechanisms into CNN has been shown to be an effective way to improve multilabel image classification. Recently, some methods have been proposed to apply attention mechanisms in remote sensing image classification.

Tong et al. proposed a channel-attention-based DenseNet (CAD) network for scene classification (Tong et al. 2020). Li et al. proposed an attention mechanism-based CNN with multi-augmented schemes (Li et al. 2020). Yu et al. proposed a feature fusion framework based on hierarchical attention and bilinear pooling for the scene classification of remote sensing images (Donghang et al. 2020).

In the MLRSIC task, there are the following studies that use label correlation to improve network performance (Liu, Shao, and Hoffmann 2021; Ni et al. 2020; Ouyang et al. 2022; Zhou et al. 2023). In the CNN-RNN framework proposed by Wang et al., the RNN model learns joint image-label embeddings from CNN features and uses the memory mechanism of RNN to predict labels in an ordered prediction path (Wang et al. 2016). Hua et al. proposed a method to extract image features through the CNN backbone network, then connect and send the image features to a bidirectional LSTM for classification, and implicitly learn label co-occurrence information (Hua, Mou, and Xiang Zhu 2019b). Zhang et al. put the label co-occurrence matrix into two convolutional layers and two fully connected layers to learn label correlations (Zhang, Hao-Yi, and Xin-Shun 2017). Li et al. proposed a cross-modal representation learning and label graph mining-based residual multi-attentional CNN-LSTM framework, which can map both visual features and label representations into a correlated space appropriately, thus improving the LSTM predictor’s ability (Li, Chen, and Zhang 2022). Huang et al. proposed an end-to-end deep learning framework consisting of a multi-scale feature fusion module, a channel-spatial attention learning module, and a label correlation extraction module (Huang, Zheng, and Huang 2021). Bazi et al. proposed a dual-branch neural network composed of an image branch and a label branch for remote sensing image classification (Bazi 2019). Zhu et al. adopted a two-layer semantic concept to annotate multilabel RS images and used a classification branch for multilabel annotation (Zhu et al. 2020).

In the study above, multiple methods were employed to implicitly restrict the output features of CNN and enforce label relationship constraints.
2.3. GNN in remote sensing

GNN is a new hot research direction in the field of artificial intelligence. Recently, many high-performance GNN models have been developed, such as graph convolution network (GCN) (Kipf and Welling 2016), graph attention network (GAT) (Velicković et al. 2017), Graph SAmple and agreGate (GraphSAGE) (Hamilton, Ying, and Leskovec 2017), etc. When GNN is used in image recognition, some strategy is usually used to construct nodes and edges as the input of GNN. The node information in GNN can be transmitted in the graph, and the characteristics of nodes in the graph can be aggregated.

In remote sensing, there are several methods to construct a scene map. Hong et al. proposed miniGCN to infer out-of-sample data without retraining networks and improving classification performance (Hong et al. 2021). Kang et al. proposed the graph relation network (GRN) method, which uses GNN to model multilabel semantic proximity (Kang et al. 2020). Li et al. proposed a method to build a scene graph for each image and generate high-level appearance features from the image as the initial node of the graph (Li et al. 2020). Liang et al. regard the ground object as a node, use the detector to detect the object, and then define the adjacent relationship between nodes according to the spatial distance between entities (Liang, Deng, and Zeng 2020). Lin et al. constructed a concept map for the whole label set by using human knowledge in conceptnet (Speer, Chin, and Havasi 2017), and then fused the two global feature vectors generated by CNN and GNN (Lin et al. 2021). In this method, the feature elements of GCN output and CNN output are multiplied, and then output to the final classifier for scene classification. However, this explicit feature fusion is highly sensitive to noise in label features.

3. Proposed approach

In this part, we elaborate on the proposed Semantic Interleaving Global Channel Attention (SIGNA) for MLRSIC. The overall flow of the proposed method is shown in Figure 2. Firstly, we introduce the specific implementation of semantical Interleaving Encoding. Secondly, we introduce the SIGNA module that uses semantical Interleaving Encoding. Finally, we look at Multi-head SIGNA that use multiple SIGNA-Based Feature Adaptive Weighting Networks. Section from front to back, each component is an important part of the latter component.

3.1. Semantical interleaving encoding

Label relationship is obtained from the overall statistics of the data set. So, for each image instance, this label relationship is noisy. Semantic interleaving coding is proposed to reduce noise sensitivity when exploiting label relations, which is inspired by interleaved codes in error correction codes. Interleaved codes are used to transform a burst channel into a random independent error channel by encoding n pieces of original data of length m into m pieces of data of length n. It is used to increase the coding’s robustness and reliability. Semantic interleaving encoding follows a similar concept. It is performed in two steps. The details of its implementation are detailed below.
3.1.1. **Using GNN to model label co-occurrence matrix as semantic features**

A co-occurrence matrix can be obtained by counting the probability of occurrence of label pairs. Calculate the probability matrix by

$$P_{ij} = \frac{N_{ij}}{N_i}, (i, j < C), \quad (1)$$

where $N_{ij}$ represents the total number of occurrences of label $i$, $N_i$ represents the total number of images where both label $i$ and label $j$ appear together, $C$ represents the number of types of labels, and all $P_{ij}$ constitute a co-occurrence matrix.

In order to prevent the possible noise caused by the long tail effect of some co-occurrence labels, the threshold $Q$ is used to filter noisy edges, and the formula can be written as:

$$G_{ij} = \begin{cases} 
0, & \text{if } P_{ij} < Q \\
1, & \text{if } P_{ij} \geq Q 
\end{cases} \quad (2)$$

where $G$ is the filtered co-occurrence matrix.

After the initial label co-occurrence matrix $G$ is constructed, the popular GCN among GNNs is chosen to model the label co-occurrence matrix. Matrix $G$ is used as the relation node of GCN. The Glove word embedding vector of the label word vector is used as input. After GCN, the initial label relationship matrix $G \in \mathbb{R}^{C \times C}$ is encoded as semantic feature $L_s \in \mathbb{R}^{C \times D}$ ($D$ usually selects the number of channels of the feature map, $C$ is the number of categories of the data set labels), thereby dispersing the label relationship into $D$ channels.

GCN updates nodes through information propagation between nodes. The goal of a GCN is to learn a function $f(\cdot, \cdot)$ on the spatial domain of a graph $G$, whose input is the feature description $H^l \in \mathbb{R}^{n \times d}$ and the corresponding correlation matrix $G \in \mathbb{R}^{n \times n}$. $I$ is the number of layers of the GCN, $n$ is the number of nodes, and $d$ is the dimension of each node feature. In the task of modelling semantic features, $n=C$, and $d=D$ for the last layer of GCN. That is, through GCN, the final discrete label relationship feature $L_s \in \mathbb{R}^{C \times D}$ is the output of GCN $H^{l+1}$. Therefore, the node feature output learned by the GCN is denoted as $H^{l+1} \in \mathbb{R}^{n \times d}$. Each GCN layer can be represented by a nonlinear function $f$.
Similarly, the GCN part of the Label Graph Relation Encoder can be replaced by the recently proposed SAGE (Hamilton, Ying, and Leskovec 2017) and GAT (Veličković et al. 2017). The experimental section details the differences between the three GNNs for modelling.

### 3.1.2. Interleaving semantic feature and image feature
At this point, the semantic feature \( L_s \in \mathbb{R}^{C \times D} \) has been obtained, and then it will be interleaved with the visual feature. Thereby, the image global features are guided into the semantic feature space with label correlation.

As shown in Figure 3(b), given a global feature vector \( Z \) of length \( D \), in order to facilitate interleaving, a linear transformation and reshape is first passed to obtain the expanded feature \( Z_s \in \mathbb{R}^{D \times C} \).

The semantic feature matrix \( L_s \) is matrix-multiplied with the extended image global feature matrix \( Z_s \). The real domain of the result is then mapped to the [0,1] space through the softmax function. Finally, the semantic interleaving matrix \( M_s \) is obtained. The calculation formula is:

\[
M_s = \text{softmax}(\text{reshape}(ZA^T + b)L_s)
\]

where reshape adjusts the number of rows and columns of \( Z \) after the linear transformation, resulting in a matrix \( Z_s \in \mathbb{R}^{D \times C} \). \( A^T \) denotes a learnable transformation matrix, and \( b \) is a learnable bias.

![Figure 3](image-url)

**Figure 3.** Pipeline for semantic interleaving encoding. (a) The calculation process of MLa-GCN (Chen et al. 2019). (b) The calculation process of semantic interleaving encoding. \( D \) is the length of global feature vector \( Z \). \( C \) is the number of categories of the data set labels.
Each row of the $L_s$ matrix represents a feature vector, and each column of the $Z_s$ matrix represents a discrete label correlation feature vector. Each row of the $M_t$ matrix represents the distribution of an eigenvalue of the image global feature vector $Z$ in the discrete semantic feature space.

Note that the proposed semantic interleaving multiplication method differs from methods such as ML-GCN (Chen et al. 2019). In ML-GCN, matrix multiplication of $Z \in \mathbb{R}^{1 \times W}$ and $L_s^T \in \mathbb{R}^{W \times C}$ is performed to obtain the final classification $Out \in \mathbb{R}^{1 \times C}$ Figure 3(a). This is equivalent to using the output matrix $L_s$ of GCN as $C$ classifiers, explicitly using semantic features for classification. In contrast, we perform matrix multiplication of $Z_t \in \mathbb{R}^{W \times C}$ and $L_s \in \mathbb{R}^{C \times W}$ to get the semantic interleaving matrix $M_t \in \mathbb{R}^{W \times W}$. This implicitly interleaves semantic and visual features, thereby implicitly guiding image global features into a semantic feature space that is highly correlated with labels Figure 3(b). This reduces the noise sensitivity of semantic features as explicit classifiers.

### 3.2. Semantic interleaving global channel attention

The following details how to apply semantic interleaving encoding to image channel features and trigger global attention in the implicit semantic feature space.

As shown in Figure 4(b), given an image feature $i \in \mathbb{R}^{D \times W \times h}$, use Avgpool to compress it in the spatial dimension. The squeezed channel feature $z \in \mathbb{R}^{D \times 1 \times 1}$ is obtained, where $D$ represents the number of channels of the image feature, and $w$ and $h$ represent the width and height of the image feature. After using semantic interleaving encoding to encode squeezed channel features, the semantic interleaving matrix $M_t$ is obtained. A fully connected layer maps the squeezed channel features $Z$ to a vector $Z_r$ of the same shape. The weighting vector $Z_w$ is obtained by matrix multiplication of $M_t$ and $Z_r$. The value of the $n$-th row of $Z_w$ is obtained by multiplying the $n$-th row of $M_t$ by $Z_r$, which means that in the semantic feature space, the first value of $Z$ is multiplied by each value of $Z$. In this way, this eigenvalue’s similarity with other eigenvalues in the semantic feature space is obtained. Finally, this vector can adaptively weight the feature maps of the full channel.

The SIGNA module is represented by the formula as follows.

$$Z = [Z_1, Z_2, \ldots, Z_D],$$

$$Z_d = \frac{1}{H \cdot W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_d(i, j),$$

$$Z_w = SIGNA(Z) = S(Z) \cdot FC(Z)$$

where $X_d(i, j)$ represents the $(i, j)$ pixel of the $d$-th channel in the feature map. $S(\cdot)$ represents the semantically interleaving encoding algorithm. $FC(\cdot)$ represents the fully connected layer.

The same is the channel attention, but the attention mechanism of SIGNA and SE-Net (Hu, Shen, and Sun 2018) is different. SE-Net models the importance of each channel through two fully connected layers without involving label correlation (Figure 4(a)). SIGNA is to achieve global attention in the semantic feature space with
Figure 4. Comparison of SIGNA and SE-Net calculation methods.

label correlation. It calculates the similarity between each channel and other channels in the semantic feature space and obtains the importance of each channel (Figure 4(b)).

3.3. Multihead SIGNA-based feature adaptive weighting networks

In order to better utilize the characteristics of remote sensing image data sets, SIGNA is designed as a plug-in network: Multihead SIGNA-Based Feature Adaptive Weighting Networks. Based on SIGNA, this network introduces the idea of the multihead mechanism and residual structure. Multiple SIGNAs can learn different features, and the multi-head mechanism can fuse multiple features learned by SIGNA, and the residual structure can well retain the original image features while introducing feature correlation. This network can be inserted into any layer of the CNN backbone to play a role.

Figure 5 shows the pipeline for inserting the proposed network into the CNN architecture. First, the squeezed channel feature Z is obtained through a pooling operation with a global receptive field, and each feature map is compressed into a feature value. Next, connect N SIGNA modules to produce N weighting vectors. N weighting vectors are concatenated in the same dimension, and then mapped into a weighting vector $Z_w$ through a fully connected layer. This final weighting vector $Z_w$ weights each feature map through a broadcast mechanism, enhancing or suppressing feature maps that are related to the label or not related to the label. Specifically, the final weighting vector $Z_w$ is dot multiplied with the feature map output3 through the
broadcast mechanism, and then this result is added with the feature map output3. Then, we get the feature map output3w, which is the weighted feature map. In order to prevent the weighting effect from being too much, inspired by the idea of residual structure, an original feature map is added after the weighted feature map. This approach ensures that the original visual features are preserved when the semantic relationship constraint effect is not optimal.

The pipeline process is represented by the formula as follows.

\[
W = FC([\text{SIGNA}_1(X), \text{SIGNA}_2(X), \ldots, \text{SIGNA}_n(X)]),
\]

\[
Y_d = X_d + X_d \cdot W_d
\]

where \( X \) represents the feature map. \( \text{SIGNA()} \) represents the SIGNA algorithm. \( FC() \) represents the fully connected layer. \( Y_d \) represents the \( d \)-th channel of output.

4. Experiments

To verify the effectiveness of the proposed method, we run it on three publicly available multilabel remote sensing image data sets. In this section, we introduce these data sets and some experimental details, then draw some useful conclusions from the analysis.

4.1. Data sets

We conduct experiments on three multilabel remote sensing data sets: AID, UCM, and DFC15 data sets. These three data sets are described in detail below.
(1) UCM Multilabel Data Set (Chaudhuri et al. 2018): Each image in the UCM multilabel data set is $256 \times 256$ pixels and contains a total of 17 classes: airplane, bare soil, buildings, cars, chaparral, court, dock, field, grass, mobile home, pavement, sand, sea, ship, tanks, trees, and water. Figure 6 displays some multilabel examples.

(2) AID Multilabel Data Set (Hua, Mou, and Xiang Zhu 2019a): The AID multilabel data has a total of 3000 multilabel images, divided into 30 scene categories, and the image size is $600 \times 600$ pixels. Similarly, the newly defined labels of the AID data set also contain 17 categories, which are consistent with the labels of the UCM data set mentioned above. Figure 7 illustrates some multilabel examples from this data set.

(3) DFC15 Multilabel Data Set (Hua, Mou, and Xiang Zhu 2019a): This data set is created from the well-known DFC15 data set. It has a total of 3342 images, each of which is $256 \times 256$ pixels, and contains a total of 8 label categories: building, boat, car, clutter, impervious, water, vegetation and trees. Some examples of multilabel annotations are given in Figure 8.

Figure 9 illustrates the distribution of the number of images in each class across the three datasets. Figure 10 visualizes the label co-occurrence matrix in the three datasets. Lastly, Figure 11 demonstrates the filtered label co-occurrence matrix in the three datasets, where noise edges have been removed, providing a clearer representation of the label relationships.

The AID and UCM data sets use the same label set, so some conclusions can be drawn by comparing the two data sets. The simpler the co-occurrence graph, the clearer and more effective the constraints between the labels. As shown in Figure 11, compared with the AID data set, the co-occurrence graph of the UCM data set has fewer and more obvious edges. This indicates that the co-occurrence relationship of the AID data set is

---

**Figure 6.** Examples of the UCM multilabel data set. (a) Soil, buildings, cars, pavement. (b) Soil, chaparral, pavement, trees. (c) Dock, ship, water. (d) Court, grass, trees. (e) Field. (f) Sand, sea. (g) Cars, grass, mobile-home, pavement. (h) Soil, buildings, cars, pavement, tanks.

**Figure 7.** Examples of the AID multilabel data set. (a) Airplane, buildings, cars, grass, pavement, trees. (b) Chaparral, sand, sea. (c) Grass, pavement, ship, water. (d) Soil, buildings, fields, trees. (e) Soil, buildings, grass, mobile-home, pavement, trees, water. (f) Buildings, cars, grass, pavement, trees. (g) Buildings, cars, courts, docks, grass, pavement, trees, water. (h) Buildings, pavement, tanks.
complex and fragmentary, while the co-occurrence relationship of the UCM data set is more obvious. Therefore, it can be speculated that the label relation method may work better on the UCM data set than on the AID data set. Section 5.3.1 demonstrates this point.

### 4.2. Evaluation metrics and loss functions

In our experiments, the performance of different models is evaluated quantitatively using the example-based F1 scores (Zhang and Zhou 2013) as evaluation metrics. To perform
a comprehensive evaluation of the network, the classification performance of each category is evaluated using label-based F1, thus analysing the impact of the network on each label. Specifically, the example-based F1 score is calculated as follows.

\[
F_e = \frac{1}{2} \frac{P_e R_e}{P_e + R_e} 
\]

where \(P_e\) and \(R_e\) are the example-based precision and recall, respectively. \(T_P_e\) represents the number of correctly predicted positive labels in an example. \(F_P_e\) represents the number of unrecognized positive labels in an example. \(F_N_e\) represents the number of mispredicted negative labels in an example. \(n\) represents the number of data set images.

The label-based F1 score are calculated as follows.

\[
F_l = \frac{1}{2} \frac{P_l R_l}{P_l + R_l} 
\]

\[
P_l = \frac{1}{c} \sum_{i=1}^{c} TP_l, \quad R_l = \frac{1}{c} \sum_{i=1}^{c} TP_l 
\]

where \(P_l\) and \(R_l\) are the label-based precision and recall, respectively. \(T_P_l\) represents the number of all correctly predicted positive labels in a class. \(F_P_l\) represents the number of unrecognized positive labels in a class. \(F_N_l\) represents the number of negative labels for all mispredictions in a class. \(c\) represents the number of data set classes.

We connect a sigmoid activation function at the end of the network and treat each value of the output layer as a binomial distribution so that the multilabel problem can be successfully transformed into a binomial classification problem on each label. We use the multilabel binary cross-entropy loss (BCE) function as the loss function. The calculation formula of the sigmoid activation function and BCE loss function (Stivaktakis, Tsagkatakis, and Tsakalides 2019; Zeggada, Melgani, and Bazi 2017) are as follows.
\[
\sigma(x) = \frac{1}{1 + \exp(-x)}
\]  

\[
\text{Loss} = - \sum_{i=1}^{n} \sum_{j=1}^{c} [\sigma(z_{ij}) \log y_{ij} + (1 - \sigma(z_{ij})) \log (1 - y_{ij})]
\]

where \( \sigma(x) \) is sigmoid activation function, \( c \) is the number of labels, \( n \) is the number of examples, \( z_{ij} \) is the output of network, and \( y_{ij} \) is the ground truth one-hot label.

### 4.3. Training details

We used three data enhancement strategies in the training process: Random Horizontal Flip, Random Vertical Flip, and Color Jitter (brightness = contrast = saturation = (0.6, 1.4)). The model is trained using the Adam optimizer, and the initial weights of the CNN backbone are set to the weights of the CNN model pre-trained on the generic data set. The initial learning rate is set to 0.001. Every 25 epochs decays by 0.1. The batch size is set to 16, and we train the model for 80 epochs.

We randomly selected 70% of the samples from each of the three data sets as the train set, 10% as the validation set, and finally 20% as the test set for training and testing respectively. The GNN in SIGNA is set to two layers, and the output dimensions are the number of channels/2 and the number of channels, respectively; The input to the GNN is the 300-dimensional glove word embedding vector for each label. All experiments are based on the same hardware and software conditions as follows: GPU: GeForce GTX 3090; OS: Ubuntu 18.04.3 LTS; CUDA Version: 11.1.0; PyTorch Version: 1.9.1 for cu111; Torch-Geometric 2.0.3 for cu111 and TorchVision Version: 0.10.1 for cu111.

### 5. Discussion

#### 5.1. Comparison of different settings

As shown in section 3, the SIGNA networks can select any layer of the backbone network to insert for global channel attention. The GNN component of SIGNA can employ popular methods such as GCN, GraphSAGE, or GAT for feature extraction. Additionally, SIGNA allows for the customization of the number of different headers to utilize the multihead mechanism. Below we will conduct experiments on the backbone networks ResNet-18 and ResNet-50 to discuss these combination options and try to find the best combination.

##### 5.1.1. Performance with different number of heads

We conducted experiments using 1, 2, 4, 6 and 8 heads, each representing one SIGNA module, with channel features replicated N times into N SIGNA module. As shown in Table 1, there is a little difference in performance for choosing different number of heads. The F1 scores of the 6-heads SIGNA module are optimal for 5 in two backbone networks and three data sets, so the multi-heads number is set as 6.
5.1.2. Selecting the performance of the different layers of CNN

SIGNA can select any layer of the backbone network for channel attention aggregation. We choose layer1, layer2, layer3 and layer4 of ResNet-18 and ResNet-50 for our experiments. The output shape of each layer of ResNet-18 is (64, 64, 64), (128, 32, 32), (256, 16, 16), (512, 8, 8). Moreover, the output shape feature maps of each layer of ResNet-50 are (256, 56, 56), (512, 28, 28), (1024, 14, 14), (2048, 7, 7). Note that we did not choose to use SIGNA in the original image. Since the original image only has 3 channels, there is no point in using channel attention on this. The shallowest layer we chose to insert is the layer1.

As shown in Table 2, AID&ResNet-18, DFC15&ResNet-18 and UCM&ResNet-50 in the experiment achieved the best performance when SIGNA was inserted in the second layer; UCM&ResNet-18 and DFC15&ResNet-50 achieved the best performance when inserted in the first layer. This indicates that SIGNA may work better in the shallow layer of CNN. This may be because, in the shallow layer of the backbone network, the number of channels is small, and the size of the feature map is large, which mainly contains the local detailed features of the image. In the deep layer of the backbone network, the number of channels is large, and the size of the feature map is small, which mainly contains abstract semantic information. Many labels of remote sensing images are shallow information, such as ocean texture, field texture, etc. There are many labels of small-sized objects in remote sensing images, such as cars, so larger feature maps work better for the proposed SIGNA.

### Table 1. Compare the number of multiheads in the SIGNA networks.

| Heads | UCM   |       |       | AID   |       |       | DFC15 |       |       |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|       | F1    | P     | R     | F1    | P     | R     | F1    | P     | R     |
| resnet18 | 1     | 93.00 | 91.73 | 94.31 | 92.18 | 92.33 | 92.03 | 96.81 | 96.91 | 96.71 |
|        | 2     | 92.85 | 92.19 | 93.51 | 92.38 | 92.85 | 91.92 | 96.63 | 96.70 | 96.56 |
|        | 4     | 92.87 | 92.07 | 93.68 | 92.41 | 92.86 | 91.97 | 96.81 | 97.03 | 96.59 |
|        | 6     | 92.93 | 92.77 | 93.10 | 92.71 | 93.25 | 92.17 | 96.96 | 97.03 | 96.90 |
|        | 8     | 92.37 | 91.17 | 93.60 | 92.31 | 92.78 | 91.84 | 96.98 | 97.32 | 96.64 |
| resnet50 | 1     | 93.78 | 93.93 | 93.64 | 92.91 | 93.18 | 92.65 | 97.33 | 97.89 | 97.68 |
|        | 2     | 93.76 | 93.71 | 93.81 | 92.69 | 93.28 | 92.11 | 97.35 | 97.77 | 96.94 |
|        | 4     | 93.84 | 94.02 | 93.66 | 92.83 | 93.87 | 91.81 | 97.24 | 97.49 | 97.00 |
|        | 6     | 93.92 | 93.84 | 94.01 | 92.93 | 93.03 | 92.83 | 97.37 | 97.86 | 96.88 |
|        | 8     | 93.87 | 93.64 | 94.10 | 92.88 | 93.33 | 92.43 | 97.34 | 97.85 | 96.83 |

The bold in the table indicates the best performance on the resnet18 and resnet50.

### 5.1.3. Performance using different GNN

In the semantic interleaving encoder, we chose the current widely used GCN, GraphSAGE and GAT, which perform excellently as relational feature extractors for our experiments. As shown in Table 3, in the experiments, four combinations achieve the best performance when
Table 3. Compare the performance of using different GNNs.

| GNN Type | UCM   | AID   | DFC15 |
|----------|-------|-------|-------|
|          | F1    | P     | R     | F1    | P     | R     | F1    | P     | R     |
| resnet18 | GCN   | 92.93 | 92.77 | 93.10 | 92.71 | 93.25 | 92.17 | 96.96 | 97.03 | 96.90 |
|          | SAGE  | 92.90 | 91.76 | 94.07 | 92.56 | 93.19 | 91.93 | 97.12 | 96.90 | 97.34 |
|          | GAT   | 92.83 | 91.48 | 94.22 | 92.69 | 93.29 | 92.09 | 96.96 | 97.33 | 96.59 |
| resnet50 | GCN   | 93.92 | 93.84 | 94.01 | 92.93 | 93.03 | 92.83 | 97.57 | 98.06 | 97.08 |
|          | SAGE  | 93.96 | 94.00 | 93.93 | 92.96 | 93.46 | 92.46 | 97.46 | 98.12 | 96.81 |
|          | GAT   | 93.91 | 93.83 | 94.00 | 92.70 | 93.45 | 91.96 | 97.40 | 97.98 | 96.83 |

The bold in the table indicates the best performance on the resnet18 and resnet50.

GraphSAGE is used in SIGNA. Therefore, we choose GraphSAGE to extract relational features in the semantic Interleaving encoder.

5.2. Comparisons with other methods and Baselines

For the comprehensive evaluation, we compare the proposed method with the following state-of-the-art multilabel classification methods. Because most of the methods are tested on the backbone networks of VGG16 and resnet50, we also use these two models as the backbone networks for experiments to compare.

- GRN (Kang et al. 2020): This method uses GCN to capture the relationships between samples, thereby enhancing the discriminative capability of the CNN.
- SE-NET (Hu, Shen, and Sun 2018): This method dynamically recalibrates the responses of individual channels by explicitly modelling the interdependencies among them.
- ML-GCN (Chen et al. 2019): The method uses GCN for label correlation extraction, which is explicitly used as a classifier for the final output of the CNN.
- SR-Net (Tan et al. 2022): The method identifies the semantic attentional regions within the extracted features and generates a category representation that is both discriminative and content-aware.
- AL-RN (Hua, Mou, and Zhu 2020): The method extracts label-specific features, identifies discriminative regions within these features, and employs a MLP layer to generate label relations for the purpose of final classification.
- CA-BiLSTM (Hua, Mou, and Zhu 2019b): This method incorporates an attention learning layer to capture class-specific features and utilizes a bidirectional LSTM network to model the relationship among classes.
- ML-CG (Lin et al. 2021): This method leverages ConceptNet to infer label correlations and combines semantic attention and label attention within GCN.
- MSGM (Lin and Chen 2022): This method directly extracts label correlations using the GCN module and simultaneously learns visual features through multi-grained semantic grouping mechanisms.
- ML-KNN (Zhang and Zhou 2007): This method uses VGG16 for feature extraction and ML-KNN for classification.
- MLRSSC-CNN-GNN (Li et al. 2020): This method constructs a scene graph for each scene and introduces a multi-layer-integration GAT model to comprehensively explore the spatio-topological relationships within the scene graph.
• Gardner (Gardner and Nichols 2017): This method uses a number of data augmentation and ensemble techniques.
• RBFNN (Zeggada, Melgani, and Bazi 2017): This method uses the VGG16 for feature extraction and the RBFNN for classification.
• Stivaktakis (Stivaktakis, Tsagkatakis, and Tsakalides 2019): This method employs a data augmentation technique that effectively expands the size of smaller datasets to significantly larger volumes.
• CNN-RNN (Wang et al. 2016): This method trains a joint image-label embedding to capture both the semantic label dependency and the relevance between images and labels.
• (Huang, Zheng, and Huang 2021): This method combines and refines the multiscale features from various layers of a CNN model by employing a channel attention mechanism. Subsequently, the label correlation information is integrated into the multiscale attentive features.
• (Zhu et al. 2020): This method uses dual-level semantic concepts, where scene labels are used to guide multilabel classification.

As shown in Table 4, the proposed method substantially improves over baseline and surpasses the current advanced multilabel classification methods in remote sensing. On the backbone of ResNet-50, the F1 score of the UCM data set improves by 12.56% and reaches 93.96%, the F1 score of the AID data set improves by 5.52% and reaches 92.96%, and the F1 score of the DFC15 data set improves by 4.33% and reaches 97.46%. Compared with ML-GCN, which also uses GCN for label-relational feature extraction, our method achieves 3.60% and 3.38% higher F1 scores on the UCM and AID data sets, respectively. This illustrates the better fusion of semantic relationship features and image features in the proposed SIGNA.

In order to further evaluate the performance of the proposed SIGNA, we study its model efficiency. Specifically, we study the computational complexity and parameters of our model. Under the same input image size and backbone network (Resnet-50), we calculate the parameters and floating point operations (FLOPs) of the proposed SIGNA method model, baseline method model, SENET method model, and ML-GCN model. As shown in Table 5, compared with the baseline method model, the proposed SIGNA model only increases 3.8% in parameters and 0.18% in FLOPs.

5.3. Effectiveness analysis

In Section 5.3.1, we analyse the influence of the complexity of the label graph on the effectiveness of SIGNA. In Section 5.3.2, we observe the impact of SIGNA networks on global features by visualizing feature embedding using t-SNE. In Section 5.3.3, we generate heatmaps for some typical examples through layerCAM (Jiang et al. 2021), analyse where the attention is focused after using SIGNA, and the impact of semantic relationships on classification.

5.3.1. Per-class case studies

As shown in Tables 6, 7 and 8, the performance of the SIGNA and baseline on the UCM, AID and DFC15 data sets for each category is shown by F1, P and R scores, respectively. In the following, we will analyse the differences in performance of each category on each data set in detail in conjunction with the label co-occurrence graphs (Figure 11).
Table 4. F1 scores of proposed method, baseline and compared SOTA methods on UCM, AID and DFC15 data sets.

| Backbone     | Method                                | UCM          | AID          | DFC15         |
|--------------|----------------------------------------|--------------|--------------|---------------|
|              |                                        | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R |
| ResNet-50    | Ours                                   | 93.96 | 94.00 | 93.93 | 92.96 | 93.46 | 92.46 | 97.46 | 98.12 | 96.81 |
|              | Baseline                               | 81.4 | 80.86 | 81.95 | 87.44 | 89.31 | 85.65 | 93.13 | 95.74 | 90.65 |
|              | GRN (Kang et al. 2020)                 | 92.4 | 91.98 | 92.83 | 91.93 | 92.79 | 91.08 | 96.24 | 96.53 | 95.95 |
|              | SE-NET (Hu, Shen, and Sun 2018)        | 89.43 | 90.94 | 87.96 | 88.91 | 90.38 | 87.48 | 94.42 | 95.56 | 93.3 |
|              | ML-GCN (Kang et al. 2020)              | 90.36 | 90.03 | 90.7 | 89.58 | 89.69 | 89.48 | - | - | - |
|              | SR-Net (Tan et al. 2022)               | 90.36 | 90.03 | 90.7 | 89.58 | 89.69 | 89.48 | - | - | - |
|              | AL-RN (Hua, Mou, and Xiang Zhu 2020)   | 88.67 | 87.96 | 89.4 | 89.97 | 89.42 | 90.52 | - | - | - |
|              | CA-BILSTM (Hua, Mou, and Xiang Zhu 2019b) | 87.93 | 88.81 | 87.07 | 89.96 | 91 | 88.95 | - | - | - |
|              | ML-CG (Lin et al. 2021)                | 83.11 | 77.94 | 89.02 | 89.01 | 89.03 | 88.99 | - | - | - |
|              | MSGM (Lin and Chen 2022)               | 85.42 | 81.34 | 89.94 | - | - | - | 95.88 | 95.68 | 96.09 |
|              | GRN (Kang et al. 2020)                 | 84.66 | 83.86 | 85.48 | - | - | - | 93.65 | 94.61 | 92.71 |
| VGG16        | Ours                                   | 92.97 | 92.51 | 93.43 | 91.78 | 91.93 | 91.63 | 96.97 | 97.37 | 96.58 |
|              | Baseline                               | 80.65 | 79.06 | 82.3 | 86.86 | 87.41 | 86.32 | 93.51 | 93.61 | 93.42 |
|              | ML-KNN (Zhang and Zhou 2007)           | 87.48 | 86.82 | 88.16 | 84.23 | 83.82 | 84.65 | - | - | - |
|              | MLRSSC-CNN-GNN Li et al. (2020b)       | 87.76 | 87.11 | 88.41 | 90.01 | 89.83 | 90.2 | - | - | - |
|              | Gardner (Gardner and Nichols 2017)     | 86.36 | 88.29 | 84.51 | 84.26 | 86.15 | 82.46 | - | - | - |
|              | RBFNN (Zeggada, Melgani, and Bazi 2017) | 88.81 | 88.87 | 88.75 | 87.53 | 88.52 | 86.56 | - | - | - |
|              | Stivaktakis (Stivaktakis, Tsagkatakis, and Tsakalides 2019) | 86.29 | 85.16 | 87.45 | 87.8 | 87.69 | 87.92 | - | - | - |
|              | CNN-RNN (Wang et al. 2016)             | 77.09 | 74.49 | 79.88 | 84.53 | 84.06 | 85.01 | - | - | - |
|              | Huang (Huang, Zheng, and Huang 2021)   | 91.74 | 90.54 | 92.98 | 91.45 | 91.03 | 91.88 | - | - | - |
| Inception-   | Zhu                                     | 91.7 | 91.75 | 91.65 | 89.06 | 89.72 | 88.41 | - | - | - |
| ResNet-v2    |                                        |              |              |              |          |          |          |          |          |          |

The bold in the table indicates the best performance on the resnet50 and vgg16.

Table 5. Comparison between the proposed method model and some methods model in parameters and FLOPs.

| Method       | Parameters | FLOPs   |
|--------------|------------|---------|
| ResNet-50    | 25,557,032 | 5.37GFlops |
| SIGNA        | 26,530,856 | 5.39GFlops |
| SE-NET       | 25,819,176 | 5.38GFlops |
| ML-GCN       | 25,710,632 | 5.38GFlops |

Table 7 illustrates the performance of various classes of the UCM data set. Tanks, courts, cars, soils and buildings have the largest improvement in F1 scores compared to baseline, with 16.67%, 14.63%, 9.77%, 9.31% and 9.07% improvement, respectively. Correspondingly, there are 3, 5, 9, 10 and 10 edges on their label co-occurrence graphs (Figure 11), respectively. In the UCM data set, cars, soil and buildings improved by 9.77%, 9.31% and 9.07%, respectively, while they only improved by only 0.22%, 2.60% and 1.59%. In the AID data set. Corresponding to these three labels, the number of edges in the AID label co-occurrence graph is 17 for cars, 19 for soil and 20 for buildings. While court (23.09%), chaparral (20.03%) and tanks (17.56%), which have the largest F1 score
### Table 6. F1, P and R scores of proposed method and baseline on each category of the DFC15 data set.

| Category        | Ours     | Baseline  |
|-----------------|----------|-----------|
|                 | F1       | P         | R         | F1       | P         | R         |
| impervious      | 99.21    | 99.40     | 99.01     | 98.61    | 98.81     | 98.42     |
| water           | 97.20    | 97.89     | 96.53     | 90.17    | 88.08     | 92.36     |
| clutter         | 95.91    | 97.67     | 94.21     | 91.33    | 93.00     | 89.71     |
| vegetation      | 95.58    | 98.78     | 92.57     | 91.22    | 90.45     | 92.00     |
| building        | 94.81    | 96.69     | 92.99     | 87.30    | 85.89     | 89.17     |
| tree            | 88.64    | 90.70     | 86.67     | 79.01    | 88.89     | 71.11     |
| boat            | 98.82    | 100.00    | 97.67     | 78.57    | 80.49     | 76.74     |
| car             | 95.28    | 95.28     | 95.28     | 76.42    | 76.42     | 76.42     |

The bold in the table shows the best performance on each category.

### Table 7. F1, P and R scores of proposed method and baseline on each category of the UCM data set.

| Category         | Ours     | Baseline  |
|------------------|----------|-----------|
|                 | F1       | P         | R         | F1       | P         | R         |
| airplane         | 100.00   | 100.00    | 100.00    | 97.44    | 100.00    | 95.00     |
| soil             | 82.78    | 81.70     | 83.89     | 73.47    | 74.48     | 72.48     |
| buildings        | 94.03    | 92.65     | 95.45     | 84.97    | 74.71     | 98.48     |
| cars             | 90.64    | 92.81     | 88.57     | 80.88    | 89.58     | 73.71     |
| chaparral        | 97.67    | 100.00    | 95.45     | 95.45    | 95.45     | 95.45     |
| court            | 97.56    | 100.00    | 95.24     | 82.93    | 85.00     | 80.95     |
| dock             | 100.00   | 100.00    | 100.00    | 97.44    | 100.00    | 95.00     |
| field            | 100.00   | 100.00    | 100.00    | 100.00   | 100.00    | 100.00    |
| grass            | 90.27    | 88.83     | 91.76     | 84.73    | 76.79     | 94.51     |
| mobile-home      | 100.00   | 100.00    | 100.00    | 95.00    | 95.00     | 95.00     |
| pavement        | 92.69    | 91.63     | 93.77     | 92.53    | 86.99     | 98.83     |
| sand             | 92.31    | 90.00     | 94.74     | 85.22    | 84.48     | 85.96     |
| sea              | 100.00   | 100.00    | 100.00    | 100.00   | 100.00    | 100.00    |
| ship             | 97.56    | 100.00    | 95.24     | 97.56    | 100.00    | 95.24     |
| tanks            | 100.00   | 100.00    | 100.00    | 83.33    | 93.75     | 75.00     |
| trees            | 91.99    | 92.23     | 91.75     | 85.45    | 78.45     | 93.81     |
| water            | 98.77    | 97.56     | 100.00    | 97.50    | 97.50     | 97.50     |

The bold in the table indicates the best performance on the resnet50 and vgg16.

### Table 8. F1, P and R scores of proposed method and baseline on each category of the AID data set.

| Category        | Ours     | Baseline  |
|-----------------|----------|-----------|
|                 | F1       | P         | R         | F1       | P         | R         |
| airplane        | 94.74    | 94.74     | 94.74     | 85.71    | 93.75     | 78.95     |
| soil            | 85.10    | 85.10     | 85.10     | 82.50    | 76.34     | 89.74     |
| buildings       | 96.37    | 96.71     | 96.03     | 94.77    | 94.23     | 95.33     |
| cars            | 94.29    | 93.49     | 95.10     | 94.07    | 94.78     | 93.38     |
| chaparral       | 51.28    | 52.63     | 50.00     | 31.25    | 41.67     | 25.00     |
| court           | 76.42    | 83.93     | 70.15     | 53.33    | 60.38     | 47.76     |
| dock            | 80.41    | 81.25     | 79.59     | 75.27    | 79.55     | 71.43     |
| field           | 84.06    | 90.63     | 78.38     | 67.61    | 70.59     | 64.86     |
| grass           | 95.38    | 93.47     | 97.37     | 93.87    | 89.15     | 99.12     |
| mobile-home     | 0.00     | 0.00      | 0.00      | 0.00     | 0.00      | 0.00      |
| pavement        | 98.06    | 98.48     | 97.63     | 96.62    | 98.01     | 95.27     |
| sand            | 91.59    | 90.74     | 92.45     | 91.26    | 94.00     | 88.68     |
| sea             | 98.82    | 100.00    | 97.67     | 83.12    | 94.12     | 74.42     |
| ship            | 76.77    | 76.00     | 77.55     | 62.22    | 68.29     | 57.14     |
| tanks           | 97.56    | 100.00    | 95.24     | 80.00    | 68.97     | 95.24     |
| trees           | 96.40    | 94.94     | 97.91     | 94.24    | 92.90     | 95.62     |
| water           | 77.38    | 87.41     | 69.41     | 67.57    | 79.37     | 58.82     |

The bold in the table indicates the best performance on the resnet50 and vgg16.
improvement in the AID data set, have 6, 5 and 6 edges on the AID label co-occurrence graphs, respectively.

The relationship between the F1 score improvement and the number of label graph edges for the above two data sets also reaffirms our previous view. The smaller the number of edges in the label co-occurrence graph, the more obvious the relationship between labels, and the larger the F1 score improvement using the SIGNA. This property is not only for the whole data set but also for each category of the data set. It is also clear from the comparison between the two data sets that the semantic relationship feature indeed brings about the F1 score improvement.

5.3.2. Visualizing feature embedding using t-SNE

T-SNE (Van der Maaten and Hinton 2008) is a feature dimensionality reduction method that reduces high-dimensional features to two-dimensional so that the feature structure can be visually represented in the figure. As shown in Figure 12, on the UCM data set, we use t-SNE to visualize the last layer feature maps of the baseline method and SIGNA. Figure 12 shows the feature distributions of the four categories of airplanes, buildings, cars, and trees. Subfigures (a), (b), (c) and (d) show the feature distributions of the SIGNA, and subfigures (e), (f), (g) and (h) show the feature distributions of the baseline method.

Globally, the feature distribution figure shows that the features of SIGNA are separated, while the features of the baseline method are not. The difference becomes more obvious when the feature points of the four categories are labelled with different colours. For SIANA, the features of the same category are highly aggregated, and the features of different categories are separated. For the baseline method, the features of the same category are highly dispersed, and different features are mixed. This indicates that after SIGNA, the channels of the feature map are weighted by semantic relations, and the image features are guided to a feature space that is highly correlated with the semantics of the labels. The feature space of the baseline method is too far from the label semantics, so the feature distribution figure after t-SNE dimensionality reduction is confusing. Observing subfigures (a) and (c), in (a) airplane category feature points are in the top pile, and in (c) cars category also has most of the feature points in the top pile. Observing subfigures (c) and (d), the trees category feature points in (d) are in the bottom left pile, while the cars category in (c) also has most of them in the bottom left pile. This illustrates

![Figure 12](image)

**Figure 12.** Visualizing feature of SIGNA and baseline method using t-SNE on the UCM data set.
that the feature aggregation is still good in the metric space constructed by the SIGNA, even though each image has multiple labels.

Notice that the Baseline method’s t-SHE figures look much worse than the SIGNA method’s t-SHE figures in Figure 12. The characteristics of t-SHE may cause this. T-SNE is a dimensionality reduction algorithm. The data points in the high-dimensional space may be very complex in the low-dimensional space, and it is difficult for t-SNE to maintain such a complex similarity relationship in dimensionality reduction. Figure 12 shows that through the guidance of the SIGNA method, in the SIGNA feature space, the features between different categories become very distinct, even after dimensionality reduction can be significantly distinguished. This facilitates multilabel classification.

5.3.3. Typical case analysis in context information extraction
LayerCAM is used to generate heatmaps, whereby the attention regions of SIGNA can be observed. LayerCAM collects object localization information from coarse to fine levels and integrates them into a high-quality class activation map where the pixels can associated with the object are better highlighted. With layerCAM, more can be observed when acting on shallow feature maps, and fine-grained details of the target object can be effectively preserved. Confidence greater than 0.5 in Figure 13 means true. The label exists in the image. Confidence less than 0.5 means false, and the label does not exist in the image. As shown, we selected two images from the UCM data set and used layerCAM to map the first layer of features in the third layer of SIGNA and the baseline method. The reason for choosing this layer for heat map plotting is that the SIGNA is inserted into the last layer of the baseline method, the second layer of the feature map, so that this layer can effectively observe the effectiveness of the SIGNA, what regions are of more interest, and whether label correlation can be exploited.

By observation, we can learn that, first, in (a) and (b), the attention distribution of ResNet-50 is similar for each classification, but the attention distribution of SIGNA differs more. Second, after passing SIGNA, the network can pay more attention to small object features, such as cars and trees in (a) and cars and tanks in (b). Compared to the baseline method, the network focuses the otherwise distracting attention on a tiny part, such as the car and tank object parts. This also illustrates the ability of SIGNA to stimulate or inhibit the correct or incorrect channel to keep their attention tightly locked on the correct object.

In subgraph (a), in the baseline method, the classification of the court is wrong, while in SIGNA, the classification of the court is correct. In the label co-occurrence map of the UCM data set, the points connected to the court are buildings, cars, grass, pavement and trees. Moreover, the baseline method correctly classified cars, grass, pavement and trees. We can speculate that through SIGNA, the network knows that all four points associated with the court exist, inferring that the court may also exist with high probability. Thus, the channels associated with the court are weighted. In subgraph (b), the classification of tanks and cars is wrong in the baseline method, while in SIGNA, the classification of tanks and cars are correct. In the label graph of the UCM data set, the points connected with tanks are buildings, soils and pavements. Moreover, the classification of cars and pavements by baseline method is correct. We can speculate that through SIGNA, the network knows that two points associated with tanks exist, inferring that tanks may also exist with high probability. Thus, the channels associated with the tanks are weighted.
6. Conclusions

This paper proposes a label correlation-based channel attention mechanism for MLRSIC. The method utilizes GNN to generate relational features between labels, and then uses semantic interleaving coding to guide image features into the feature space related to semantic relations. Global channel attention is implemented in this feature space, in which the features of the same category in this space are highly aggregated, and the features of
different categories are obviously separated. SIGNA uses a simple label co-occurrence relationship to weight channel features for better classification accuracy. Experimental results on three multilabel data sets, UCM, AID and DFC15, show that the proposed SIGNA exhibits the best performance compared to popular multilabel image classification methods.

More ways to encode label relationships could be explored in future works. In this paper, the label co-occurrence matrix is used to construct the graph explicitly. The Glove word embedding of the labels is used as the input of the GNN. The correlation between word embedding and original labels is weak. A better encoding that closely related to the original labels could be explored. And multilevel feature fusion methods could be explored in future work. The proposed SIGNA blocks can be inserted into any layer of CNN. Inserting SIGNA in several layers may result in better fusion features than only in one shallow layer. The co-occurrence graph of input GNN dynamically adjusted at runtime is also a method that can be studied in the future.

Acknowledgements

We would like to thank Qingdao University of Technology for their technical support, as well as all those who participated in this paper. This work was supported in part by the National Natural Science Foundation of China under Grant 62171247.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported in part by the National Natural Science Foundation of China under Grant 62171247.

ORCID

Lijian Zhou http://orcid.org/0000-0002-5716-0772
Kun Zhao http://orcid.org/0000-0001-9949-4693

Data availability of statement

The code and data that support the findings of this study are available from the first author, Yongkun Liu (Email: yongkunliu.mail@gmail.com), upon reasonable request. Some useful information is also available at https://github.com/kyle-one/SIGNA.

Authors contribution

All the authors made significant contributions to this work. Project administration, K.X.; Innovations and original draft writing, Y.L.; Coding, K.N.; Review and editing, Y.Z. and J.Z. All authors have read and agreed to the published version of the manuscript.
References

Alshehri, A., Y. Bazi, N. Ammour, H. Almubarak, and N. Alajlan. 2019. “Deep Attention Neural Network for Multi-Label Classification in Unmanned Aerial Vehicle Imagery.” *Institute of Electrical and Electronics Engineers Access* 7:119873–119880. https://doi.org/10.1109/ACCESS.2019.2936616.

Bazi, Y. 2019. “Two-Branch Neural Network for Learning Multi-Label Classification in UAV Imagery.” In *2019 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Yokohama, Japan, 2443–2446. https://doi.org/10.1109/igarss.2019.8898895.

Chaudhuri, B., B. Demir, S. Chaudhuri, and L. Bruzzone. 2018. “Multilabel Remote Sensing Image Retrieval Using a Semisupervised Graph-Theoretic Method.” *IEEE Transactions on Geoscience and Remote Sensing* 56 (2): 1144–1158. https://doi.org/10.1109/TGRS.2017.2760909.

Chen, Z.-M., X.-S. Wei, P. Wang, and Y. Guo. 2019. “Multi-Label Image Recognition with Graph Convolutional Networks.” In *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 5177–5186. https://doi.org/10.1109/cvpr.2019.00532.

Chenyu, L., B. Zhang, D. Hong, J. Yao, and J. Chanussot. 2023. “LRR-Net: An Interpretable Deep Unfolding Network for Hyperspectral Anomaly Detection.” *IEEE Transactions on Geoscience and Remote Sensing* 61:1–12. https://doi.org/10.1109/TGRS.2023.3279834.

Dai, O. E., B. Demir, B. Sankur, and L. Bruzzone. 2018. “A Novel System for Content-Based Retrieval of Single and Multi-Label High-Dimensional Remote Sensing Images.” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11 (7): 2473–2490. https://doi.org/10.1109/JSTARS.2018.2832985.

Donghang, Y., H. Guo, X. Qing, L. Jun, C. Zhao, and Y. Lin. 2020. “Hierarchical Attention and Bilinear Fusion for Remote Sensing Image Scene Classification.” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13:6372–6383. https://doi.org/10.1109/JSTARS.2020.3030257.

Fernandez-Beltran, R., A. Plaza, J. Plaza, and F. Pla. 2018. “Hyperspectral Unmixing Based on Dual-Depth Sparse Probabilistic Latent Semantic Analysis.” *IEEE Transactions on Geoscience & Remote Sensing* 56 (11): 6344–6360. https://doi.org/10.1109/TGRS.2018.2837150.

Fernandez-Beltran, R., A. Plaza, J. Plaza, and F. Pla. 2020. “Endmember Extraction from Hyperspectral Imagery Based on Probabilistic Tensor Moments.” *IEEE Geoscience & Remote Sensing Letters* 17 (12): 2120–2124. https://doi.org/10.1109/LGRS.2019.2963114.

Gao, L., B. Zhao, X. Jia, W. Liao, and B. Zhang. 2017. “Optimized Kernel Minimum Noise Fraction Transformation for Hyperspectral Image Classification.” *Remote Sensing* 9 (6): 548. https://doi.org/10.3390/RS9060548.

Gardner, D., and D. Nichols. 2017. *Multi-label Classification of Satellite Images with Deep Learning*. Technical Report. Palo Alto, CA: Stanford University. http://vision.stanford.edu/teaching/cs231n/reports/2017/pdfs/908.pdf.

Geng, J., J. Fan, H. Wang, M. Xiaorui, L. Baoming, and F. Chen. 2015. “High-resolution SAR image classification via deep convolutional autoencoders.” *IEEE Geoscience and Remote Sensing Letters* 12 (11): 2351–2355. https://doi.org/10.1109/LGRS.2015.2478256.

Hamilton, W. L., Z. Ying, and J. Leskovec. 2017. “Inductive Representation Learning on Large Graphs.” In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, 1024–1034. Long Beach, CA, USA.

Hong, D., L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot. 2021. “Graph Convolutional Networks for Hyperspectral Image Classification.” *IEEE Transactions on Geoscience & Remote Sensing* 59 (7): 5966–5978. https://doi.org/10.1109/TGRS.2020.3015157.

Hong, D., L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang. 2021. “More Diverse Means Better: Multimodal Deep Learning Meets Remote-Sensing Imagery Classification.” *IEEE Transactions on Geoscience & Remote Sensing* 59 (5): 4340–4354. https://doi.org/10.1109/TGRS.2020.3016820.

Hong, D., N. Yokoya, J. Chanussot, X. Jian, and X. Xiang Zhu. 2019. “Learning to Propagate Labels on Graphs: An Iterative Multitask Regression Framework for Semi-Supervised Hyperspectral Dimensionality Reduction.” *Isprs Journal of Photogrammetry & Remote Sensing* 158:35–49. https://doi.org/10.1016/j.isprsjprs.2019.09.008.
Hong, D., N. Yokoya, J. Chanussot, and X. Xiang Zhu. 2018. “An Augmented Linear Mixing Model to Address Spectral Variability for Hyperspectral Unmixing.” IEEE Transactions on Image Processing 28 (4): 1923–1938. https://doi.org/10.1109/TIP.2018.2878958.

Hong, D., N. Yokoya, G. Nan, J. Chanussot, and X. X. Zhu. 2019. “Learnable Manifold Alignment (LeMa): A Semi-Supervised Cross-Modality Learning Framework for Land Cover and Land Use Classification.” Isprs Journal of Photogrammetry & Remote Sensing 147:193–205. https://doi.org/10.1016/j.isprsjprs.2018.10.006.

Hua, Y., L. Mou, and X. Xiang Zhu. 2019a. “Label Relation Inference for Multi-Label Aerial Image Classification.” In 2019 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Yokohama, Japan, 5244–5247. https://doi.org/10.1109/igarss.2019.8898934.

Hua, Y., L. Mou, and X. Xiang Zhu. 2019b. “Recurrently Exploring Class-Wise Attention in a Hybrid Convolutional and Bidirectional LSTM Network for Multi-Label Aerial Image Classification.” Isprs Journal of Photogrammetry & Remote Sensing 149:188–199. https://doi.org/10.1016/j.isprsjprs.2019.01.015.

Hua, Y., L. Mou, and X. Xiang Zhu. 2020. “Relation Network for Multilabel Aerial Image Classification.” IEEE Transactions on Geoscience and Remote Sensing 58 (7): 4558–4572. https://doi.org/10.1109/TGRS.2019.2963364.

Huang, R., F. Zheng, and W. Huang. 2021. “Multilabel Remote Sensing Image Annotation with Multiscale Attention and Label Correlation.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14:6951–6961. https://doi.org/10.1109/JSTARS.2021.3091134.

Hu, J., L. Shen, and G. Sun. 2018. “Squeeze-and-Excitation Networks.” In 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 7132–7141. https://doi.org/10.1109/cvpr.2018.00745.

Jiang, P.-T., C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei. 2021. “LayerCam: Exploring Hierarchical Class Activation Maps for Localization.” IEEE Transactions on Image Processing 30:5875–5888. https://doi.org/10.1109/TIP.2021.3089943.

Ju, Y., J. Dong, and S. Chen. 2021. “Recovering Surface Normal and Arbitrary Images: A Dual Regression Network for Photometric Stereo.” IEEE Transactions on Image Processing 30:3676–3690. https://doi.org/10.1109/TIP.2021.3064230.

Ju, Y., Y. Peng, M. Jian, F. Gao, and J. Dong. 2022. “Learning Conditional Photometric Stereo with High-Resolution Features.” Computational Visual Media 8 (1): 105–118. https://doi.org/10.1007/S41095-021-00223-Y.

Kaiming, H., X. Zhang, S. Ren, and J. Sun. 2016. “Deep Residual Learning for Image Recognition.” In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 770–778. https://doi.org/10.1109/cvpr.2016.90.

Kang, J., R. Fernandez-Beltran, D. Hong, J. Chanussot, and A. Plaza. 2020. “Graph Relation Network: Modeling Relations Between Scenes for Multilabel Remote-Sensing Image Classification and Retrieval.” IEEE Transactions on Geoscience and Remote Sensing 59 (5): 4355–4369. https://doi.org/10.1109/TGRS.2020.3016020.

Kipf, T. N., and M. Welling. 2016. “Semi-Supervised Classification with Graph Convolutional Networks.” CoRR Abs/1609.02907. http://arxiv.org/abs/1609.02907.

Koda, S., A. Zeggada, F. Melgani, and R. Nishii. 2018. “Spatial and Structured SVM for Multilabel Image Classification.” IEEE Transactions on Geoscience and Remote Sensing 56 (10): 5948–5960. https://doi.org/10.1109/TGRS.2018.2828862.

Liang, J., Y. Deng, and D. Zeng. 2020. “A Deep Neural Network Combined CNN and GCN for Remote Sensing Scene Classification.” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13:4325–4338. https://doi.org/10.1109/JSTARS.2020.3011333.

Li, P., P. Chen, and D. Zhang. 2022. “Cross-Modal Feature Representation Learning and Label Graph Mining in a Residual Multi-Attentional CNN-LSTM Network for Multi-Label Aerial Scene Classification.” Remote Sensing 14 (10): 2424. https://doi.org/10.3390/RS14102424.

Li, Y., R. Chen, Y. Zhang, M. Zhang, and L. Chen. 2020. “Multi-Label Remote Sensing Image Scene Classification by Combining a Convolutional Neural Network and a Graph Neural Network.” Remote Sensing 12 (23): 4003. https://doi.org/10.3390/RS12234003.
Li, F., R. Feng, W. Han, and L. Wang. 2020. “An Augmentation Attention Mechanism for High-Spatial-Resolution Remote Sensing Image Scene Classification.” *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 13:3862–3878. https://doi.org/10.1109/JSTARS.2020.3006241.

Lin, T.-Y., S. Belongie, and J. Hays. 2013. “Cross-View Image Geolocalization.” In *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Portland, OR, USA, 891–898. https://doi.org/10.1109/cvpr.2013.120.

Lin, D., J. Lin, Z. J. W. Liang Zhao, Z. Chen, and Z. Chen. 2021. “Multilabel Aerial Image Classification with a Concept Attention Graph Neural Network.” *IEEE Transactions on Geoscience and Remote Sensing* 60:1–12. https://doi.org/10.1109/TGRS.2020.3041461.

Liu, Y., Z. Shao, and N. Hoffmann. 2021. “Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions.” *arXiv preprint arXiv:2112.05561*.

Ma, A., Y. Wan, Y. Zhong, J. Wang, and L. Zhang. 2021. “SceneNet: Remote Sensing Scene Classification Deep Learning Network Using Multi-Objective Neural Evolution Architecture Search.” *ISPRS Journal of Photogrammetry and Remote Sensing* 172:171–188. https://doi.org/10.1016/j.isprsjprs.2020.11.025.

Movshovitz-Attias, Y., Y. Qian, M. C. Stumpe, V. D. Shet, S. Arnoud, and L. Yatziv. 2015. “Ontological Supervision for Fine Grained Classification of Street View storefronts.” In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 1693–1702. https://doi.org/10.1109/cvpr.2015.7298778.

Ni, J., W. Jianhuang, H. Wang, J. Tong, Z. Chen, K. K. Wong, and D. Abbott. 2020. “Global Channel Attention Networks for Intracranial Vessel Segmentation.” *Computers in Biology and Medicine* 118:103639. https://doi.org/10.1016/J.COMPBIOMED.2020.103639.

Ouyang, S., X. Jiahui, W. Chen, Y. Dong, L. Xianju, and L. Jun. 2022. “A Fine-Grained Genetic Landform Classification Network Based on Multimodal Feature Extraction and Regional Geological Context.” *IEEE Transactions on Geoscience and Remote Sensing* 60:1–14. https://doi.org/10.1109/TGRS.2022.3203606.

Rasti, B., D. Hong, R. Hang, P. Ghamisi, X. Kang, J. Chanussot, and J. Atli Benediktsson. 2020. “Feature Extraction for Hyperspectral Imagery: The Evolution from Shallow to Deep: Overview and Toolbox.” *IEEE Geoscience and Remote Sensing Magazine* 8 (4): 60–88. https://doi.org/10.1109/MGRS.2020.2979764.

Simonyan, K., A. Zisserman, and A. Zisserman. 2014. “Learning Local Feature Descriptors Using Convex Optimisation.” *IEEE Transactions on Pattern Analysis & Machine Intelligence* 36 (8): 1573–1585. *arXiv preprint arXiv:1409.1536*. https://doi.org/10.1109/TPAMI.2014.2301163.

Speer, R., J. Chin, and C. Havasi. 2017. “ConceptNet 5.5: An Open Multilingual Graph of General Knowledge.” In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, California*, USA, 4444–4451. https://doi.org/10.1609/aaai.v31i1.11164.

Stivaktakis, R., G. Tsagkatakis, and P. Tsakalides. 2019. “Deep learning for multilabel land cover scene categorization using data augmentation.” *IEEE Geoscience and Remote Sensing Letters* 16 (7): 1031–1035. https://doi.org/10.1109/LGRS.2019.2893306.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. “Going Deeper with Convolutions.” In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 1–9. https://doi.org/10.1109/cvpr.2015.7298594.

Tan, Q., Y. Liu, X. Chen, and Y. Guoxian. 2017. “Multi-Label Classification Based on Low Rank Representation for Image Annotation.” *Remote Sensing* 9 (2): 109. https://doi.org/10.3390/RS9020109.

Tan, X., Z. Xiao, J. Zhu, Q. Wan, K. Wang, and L. Deren. 2022. “Transformer-Driven Semantic Relation Inference for Multilabel Classification of High-Resolution Remote Sensing Images.” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15:1884–1901. https://doi.org/10.1109/JSTARS.2022.3145042.

Tong, W., W. Chen, W. Han, L. Xianju, and L. Wang. 2020. “Channel-Attention-Based DenseNet Network for Remote Sensing Image Scene Classification.” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13:4121–4132. https://doi.org/10.1109/JSTARS.2020.3009352.

Van der Maaten, L., and G. Hinton. 2008. “Visualizing data using t-SNE.” *Journal of Machine Learning Research* 9 (11): 2579–2605.
Veličković, P., G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. 2017. “Graph Attention Networks.” arXiv preprint arXiv:1710.10903.
Vucetic, B., and J. Yuan. 2012. Turbo Codes: Principles and Applications. New York, NY: Springer.
Wang, X., X. Xiong, and C. Ning. 2019. “Multi-label remote sensing scene classification using multi-bag integration.” Institute of Electrical and Electronics Engineers Access 7:120399–120410. https://doi.org/10.1109/ACCESS.2019.2937188.
Wang, J., Y. Yang, J. Mao, Z. Huang, C. Huang, and X. Wei. 2016. “CNN-RNN: A Unified Framework for Multi-Label Image Classification.” In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)Las Vegas, NV, USA, 2285–2294. https://doi.org/10.1109/cvpr.2016.251.
Yao, J., B. Zhang, L. Chenyu, D. Hong, and J. Chanussot. 2023. “Extended Vision Transformer (ExViT) for Land Use and Land Cover Classification: A Multimodal Deep Learning Framework.” IEEE Transactions on Geoscience and Remote Sensing 61:1–15. https://doi.org/10.1109/TGRS.2023.3284671.
Zeggada, A., S. Benbraika, F. Melgani, and Z. Mokhtari. 2018. “Multilabel Conditional Random Field Classification for UAV Images.” IEEE Geoscience and Remote Sensing Letters 15 (3): 399–403. https://doi.org/10.1109/LGRS.2018.2790426.
Zeggada, A., F. Melgani, and Y. Bazi. 2017. “A Deep Learning Approach to UAV Image Multilabeling.” IEEE Geoscience and Remote Sensing Letters 14 (5): 694–698. https://doi.org/10.1109/LGRS.2017.2671922.
Zhang, P.-F., W. Hao-Yi, and X. Xin-Shun. 2017. “A Dual-CNN Model for Multi-Label Classification by Leveraging Co-Occurrence Dependencies Between Labels.” In Advances in Multimedia Information Processing–PCM 2017: 18th Pacific-Rim Conference on Multimedia, Harbin, China, September 28-29, 2017, Revised Selected Papers, Part I 18, 315–324. Springer. https://doi.org/10.1007/978-3-319-77380-3_30.
Zhang, M.-L., and Z.-H. Zhou. 2007. “ML-KNN: A Lazy Learning Approach to Multi-Label Learning.” Pattern Recognition 40 (7): 2038–2048. https://doi.org/10.1016/J.PATCOG.2006.12.019.
Zhang, M.-L., and Z.-H. Zhou. 2013. “A Review on Multi-Label Learning Algorithms.” IEEE Transactions on Knowledge and Data Engineering 26 (8): 1819–1837. https://doi.org/10.1109/TKDE.2013.39.
Zhao, K., Y. Liu, S. Hao, L. Shaoxing, H. Liu, and L. Zhou. 2021. “Bounding Boxes are All We Need: Street View Image Classification via Context Encoding of Detected Buildings.” IEEE Transactions on Geoscience and Remote Sensing 60:1–11. https://doi.org/10.1109/TGRS.2020.3042202.
Zhao, W., and D. Shihong. 2016. “Spectral–Spatial Feature Extraction for Hyperspectral Image Classification: A Dimension Reduction and Deep Learning Approach.” IEEE Transactions on Geoscience and Remote Sensing 54 (8): 4544–4554. https://doi.org/10.1109/TGRS.2016.2543748.
Zhou, G., X. Jiahui, W. Chen, L. Xianju, L. Jun, and L. Wang. 2023. “Deep Feature Enhancement Method for Land Cover with Irregular and Sparse Spatial Distribution Features: A Case Study on Open-Pit Mining.” IEEE Transactions on Geoscience and Remote Sensing 61:1–20. https://doi.org/10.1109/TGRS.2023.3336053.
Zhu, X. X., H. Jingliang, C. Qiu, Y. Shi, J. Kang, L. Mou, H. Bagheri, et al. 2019. “So2Sat LCZ42: A benchmark dataset for global local climate zones classification.” arXiv preprint arXiv:1912.12171
Zhu, P., Y. Tan, L. Zhang, Y. Wang, J. Mei, H. Liu, and W. Mengfan. 2020. “Deep Learning for Multilabel Remote Sensing Image Annotation with Dual-Level Semantic Concepts.” IEEE Transactions on Geoscience and Remote Sensing 58 (6): 4047–4060. https://doi.org/10.1109/TGRS.2019.2960466.