SAGN: Semantic-Aware Graph Network for Remote Sensing Scene Classification

Yuqun Yang, *Student Member, IEEE*, Xu Tang, *Senior Member, IEEE*, Yiu-Ming Cheung, *Fellow, IEEE*, Xiangrong Zhang, *Senior Member, IEEE*, and Licheng Jiao, *Fellow, IEEE*

***Abstract*— The scene classification of remote sensing (RS) images plays an essential role in the RS community, aiming to assign the semantics to different RS scenes. With the increase of spatial resolution of RS images, high-resolution RS (HRRS) image scene classification becomes a challenging task because the contents within HRRS images are diverse in type, various in scale, and massive in volume. Recently, deep convolution neural networks (DCNNs) provide the promising results of the HRRS scene classification. Most of them regard HRRS scene classification tasks as single-label problems. In this way, the semantics represented by the manual annotation decide the final classification results directly. Although it is feasible, the various semantics hidden in HRRS images are ignored, thus resulting in inaccurate decision. To overcome this limitation, we propose a semantic-aware graph network (SAGN) for HRRS images. SAGN consists of a dense feature pyramid network (DFPN), an adaptive semantic analysis module (ASAM), a dynamic graph feature update module, and a scene decision module (SDM). Their function is to extract the multi-scale information, mine the various semantics, exploit the unstructured relations between diverse semantics, and make the decision for HRRS scenes, respectively. Instead of transforming single-label problems into multi-label issues, our SAGN elaborates the proper methods to make full use of diverse semantics hidden in HRRS images to accomplish scene classification tasks. The extensive experiments are conducted on three popular HRRS scene data sets. Experimental results show the effectiveness of the proposed SAGN. Our source codes are available at https://github.com/TangXu-Group/SAGN.**

***Index Terms*— High resolution remote sensing image, scene classification, deep learning.**

Manuscript received 15 March 2021; revised 7 October 2021 and 6 March 2022; accepted 10 January 2023. Date of publication 24 January 2023; date of current version 31 January 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62171332 and Grant 62276197; in part by the National Key Laboratory of Science and Technology on Remote Sensing Information and Imagery Analysis, Beijing Research Institute of Uranium Geology, under Grant 6142A010301; in part by the NSFC/Research Grants Council (RGC) Joint Research Scheme under Grant N_HKBU214/21; in part by the General Research Fund of RGC under Grant 12201321 and Grant 12202622; and in part by the Hong Kong Baptist University (HKBU) under Grant RC-FNRA-IG/18-19/SCI/03. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ghassan Alregib. *(Corresponding authors: Xu Tang; Yiu-Ming Cheung.)*

Yuqun Yang, Xu Tang, Xiangrong Zhang, and Licheng Jiao are with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi’an, Shaanxi 710071, China (e-mail: tangxu128@gmail.com).

Yiu-Ming Cheung is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, SAR, China (e-mail: ymc@comp.hkbu.edu.hk).

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2023.3238310, provided by the authors.

Digital Object Identifier 10.1109/TIP.2023.3238310

I. Introduction

WITH the improvement of the resolution of remote sensing (RS) images, more detailed land-cover information can be shown in the high-resolution RS (HRRS) images. According to the different semantics of land-covers, HRRS images can be classified into different scenes. The HRRS image scene classification becomes important increasingly as it can be used in many RS applications, see [1], [2], [3]. However, it is a tough and challenging task since HRRS images are complicated in contents, diverse in semantics, multi-scaled in targets, and huge in volume. Accordingly, how to improve the classification accuracy of HRRS scenes becomes a hot research topic in the RS community.

A lot of methods have been proposed to distinguish HRRS scenes [4], [5], [6]. The two main parts, feature extractor and classifier, play the crucial role in this task. The feature extractor aims to map HRRS images into the proper visual features, while the classifier focuses on grouping HRRS scenes into different various semantic classes. Due to the favorable stability and efficiency, the hand-crafted features (e.g., texture features [7], [8], spectral features [9], [10], color features [11], [12], and shape features [13], [14]) and traditional classifiers (e.g., support vector machine [15] and decision tree [16]) are often used. However, since the hand-crafted features are hard to describe the information of HRRS image comprehensively and traditional classifiers cannot match the information distributions of hand-crafted features perfectly, the performance of HRRS image scene classification cannot meet what we expect. With the development of deep convolutional neural networks (DCNNs) [17], DCNN-based classification methods become popular increasingly. The deep features learned by the hierarchical DCNNs can describe HRRS images more completely in comparison to the hand-crafted features. Furthermore, the classifiers within DCNN-based methods are trained with feature extractors together so that they can follow the deep feature distributions appropriately. In fact, more and more DCNN-based methods have been proposed for HRRS image scene classification tasks with achieving impressive results in a variety of applications [18], [19], [20], [21], [22], [23], [24].

Generally speaking, the HRRS image scene classification is a single-label task. However, the pre-defined single label is not able to fully describe the complex contents of an HRRS scene. Let us take a “bridge” scene as an example (see Fig. 1). Besides the “bridge”, there are still some regions

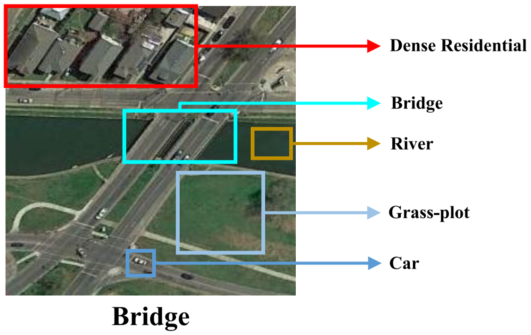


Fig. 1. The above high-resolution remote sensing (HRRS) image is selected from the UC Merced data set [25]. Its manual semantic scene label is "bridge". It can be seen that apart from the "bridge", there are many other distinct land-covers, including "river", "grass-plot", "car", and "dense residential". If we only consider the "bridge" during the feature learning, the contents corresponding to other semantics would decrease the discrimination of learned features and impact the classification negatively. Hence, the complex contents corresponding to diverse semantics should be taken into account simultaneously.

covered by "river", "car", "grass-plot", and even "dense building". Therefore, rather than only considering "bridge", exploring latent semantic contents and analyzing the context relations among them are crucial for the scene classification. The common DCNN-based methods can capture various contents but deem them equally. This leads to information disturbance which would decrease the discrimination of learned features and impact the classification negatively. To highlight important contents for the scene classification, researchers have introduced the attention mechanism into DCNN-based methods [26], [27], [28], which pushes the networks to pay more attention to the essential semantics. Although the attention mechanism can partly deal with information disturbance, the context relations among various latent information hidden in HRRS images cannot be explored.

Recently, due to the solid capacity in mining relationships, graph convolutional network (GCN) has attracted more and more scholars' attention [29], [30]. Apart from discovering diverse semantics within HRRS scenes, GCNs can exploit their intrinsic relationships at the same time. Accordingly, the obtained features are more representative. When researchers use GCNs to interpret HRRS images, instead of constructing a graph at the pixel level (which is time and storage consuming), they are accustomed to dividing an HRRS image into several regions and constructing the initial graph at the region level. Meanwhile, the original nodes' representation is the average features of their HRRS pixels. Then, the region representation can be obtained by updating the graph nodes using the graph convolution, which considers their neighborhoods' information and the intrinsic connection between them simultaneously. After this, the high-level feature of the HRRS image can be generated. The acquired high-level feature can reflect the short-range semantics (local) and long-range context information (global) of the HRRS image.

Although the regular GCNs mentioned above are available in exploring the HRRS images' complex and diverse contents, they have a distinct shortcoming, i.e., the knowledge stored in pixels may be lost when constructing the initial graph. Also, the pixels' information will not be considered during the graph convolution. The above deficiency somewhat limits GCNs' performance in HRRS scene understanding. Therefore, we propose a new graph node updating strategy based on the dynamic graph theory [31] to overcome this issue. Suppose a region-level graph has already been established for an HRRS scene. Rather than updating a node representation directly, the developed strategy renews the HRRS pixels belonging to this node by the graph convolution under the region-level adjacency matrix. Then, the node would be updated by considering all of the renewed HRRS pixels. The updating strategy discussed above and the difference between it and the traditional node updating scheme are displayed in Fig. 2. Based on this strategy, a simple yet effective RS scene classification method named semantic-aware graph network (SAGN) is proposed in this paper for HRRS scene classification. First, SAGN adopts a dense feature pyramid network (DFPN) [32] to extract representative feature maps. Second, an adaptive semantic analysis module is proposed to obtain the diverse semantics within HRRS images. By analyzing the extracted feature maps, HRRS scenes can be analyzed into various semantic regions adaptively. Third, these regions are organized in a graph, and the introduced graph node updating strategy is adopted to capture their high-level information. In this way, both the pixel-level semantics and the region-level context information within RS scenes can be captured for classification. Finally, the final scene category is decided according to the diverse semantic regions.

The main contributions of this paper are summarized as follows:

1) An end-to-end HRRS image scene classification network (SAGN) is proposed. Not only can the diverse semantics hidden in HRRS scenes be explored but also the relationships among various semantics can be captured, which is beneficial to understand the complex contents within HRRS scenes.
2) An adaptive semantic analysis module is proposed, which can analyze the feature maps of HRRS scenes to different semantic regions adaptively. Accordingly, the diverse latent semantics of HRRS scenes can be discovered.
3) A helpful graph node updating strategy based on dynamic GCN is proposed for HRRS scene understanding. The key point of our updating strategy is to renew the HRRS pixels' features using the region-level adjacency relations. In this way, the pixel-level knowledge, region-level information, and their context connections can be captured.

The rest of this paper is organized as follows. Section II makes an overview of deep-learning-based HRRS scene classification methods. In Section III, the proposed SAGN with three sub-modules is introduced in detail. The experiments and discussion are conducted in Section IV. Section V draws a brief conclusion.

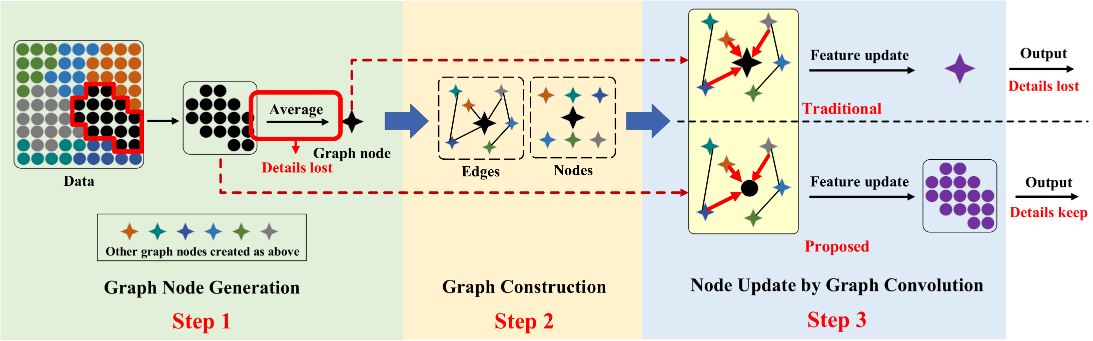


Fig. 2. Schematic illustration of our proposed and traditional graph node updating strategy. Assume that an HRRS scene has been divided into several regions. After the average and similarities calculation, the initial graph can be constructed (see Step1 and Step 2). Taking the region marked by the red frame as an example, its original corresponding graph node (black star) can be produced by averaging the pixels' features in this region (black circle). In addition, when all graph nodes have been obtained, the edges among them can be decided by their similarities. Then, the original graph nodes would be updated by the graph convolution. In the traditional updating strategy (top in Step 3), the graph nodes will be directly renewed by the graph convolution (see the transformation from black star to purple star). In contrast, the graph nodes will be updated in two steps in the proposed graph node updating strategy. The first step is to renew the features of pixels belonging to a node (region) under the region-level adjacency relations (see the transformation from black circle to purple circle). The second step is to update the node's representation by averaging the new pixel features. In this way, both the local semantics and the global context information within HRRS scenes can be captured for classification.

II. Related Work

This section will make an overview of the existing HRRS image scene classification methods by dividing them into two groups, i.e., the general and semantic-aware deep scene classification methods.

A. General Deep Scene Classification Methods

With the vigorous development of deep learning, deep-learning-based HRRS scene classification methods are dominated in the RS community and achieve impressive performance due to the strong capacity of feature learning [33].

At first, scholars are inclined to adopt the pre-trained DCNNs to complete HRRS scene classification. For example, Dimitrios et al. [34] introduced the Overfeat network [35] that is pre-trained by ImageNet data set [36] to extract visual features from HRRS images. Then, the extracted features are reshaped into the 2D matrixes and the final classification results are obtained by another shallow CNN. Although this kind of solution is feasible, the complex characteristics (e.g., scale variation of objects) of HRRS images are not fully considered. To overcome this limitation, Liu et al. [37] proposed a multi-scale convolutional neural network (MCNN) for HRRS image scene classification. MCNN constructs dual-branch nets to explore the multi-scale information within HRRS images so that the scale-invariant features can be extracted to ensure the classification results. An HRRS image scene classification method was proposed in [38], where the global features and rearranged local features are captured from HRRS images respectively so that the classification results are positive. Similar to [38], Zheng et al. [39] proposed a deep scene representation model to learn invariance CNN features from HRRS scenes, in which the geometric invariance information is supplemented to the visual features so that the discrimination of extracted features can be further improved. Meanwhile, a novel representation based on a ConvNet with context aggregation was proposed in [40], which adopts two-pathway ResNet to capture local details and regional context for improving the identifiability of features. Feature stack strategy is also employed to obtain rich features for classification, where multilayer stacked covariance pooling (MSCP) is proposed to utilize the covariance matrix to stack deep features.

Besides the mentioned deep methods which focus on learning the visual features from HRRS images directly, the feature aggregation scheme is also popular in this field [41], [42]. However, the process of aggregating features is generally unsupervised and subjective, which limits the improvement of performance [43]. Therefore, an end-to-end feature aggregation CNN (FACNN) was proposed in [43]. FACNN leverages the semantic label information to enrich the visual features so that the classification results can be improved. In [44], a local Binary Patterns (LBP) encoded CNN model was proposed. It is trained by mapped coded images with explicit LBP based texture information. The additional texture features would provide complementary information to the standard RGB deep models for better classification performance.

The above HRRS scene classification methods assume that the semantic labels are correct. However, it is well-known that HRRS scene annotation is a time-consuming and tough task. The incorrect annotation or improper semantic labels are unavoidable. Therefore, some works aim to complete the HRRS scene classification under the incorrect annotation scenario. To reduce the influence from incorrect scene annotation, an error-tolerant deep learning approach for HRRS image scene classification was proposed in [45]. It learns multiview CNNs and corrects error annotations alternatively in an iterative manner so that the negative effects of incorrect annotations can be mitigated.

B. Semantic-Aware Deep Scene Classification Methods

The methods in this group focus on mining the detailed semantic information hidden in the HRRS images for improving HRRS scene classification.

As the classical method for analyzing the semantics of images, the attention mechanism [46] is received great attention in the computer vision as well as the RS community. Wang et al. [27] proposed an end-to-end network named attention recurrent convolutional network (ARCNet), which can select key regions adaptively from high-level features to improve the discrimination of features so that the scene classification performance can be promoted. Although the self-attention mechanism used in ARCNet is effective, it is time-consuming in practical. To solve this problem, the thrifty attention and its recurrent version were proposed [28], which can collect the global contextual information with low computational complexity. Also, as a general module, they can be embedded in any CNN model. Tang et al. [47] introduced a dual-channel network to address HRRS scene classification. The proposed model combines the spatial and spectral attention to capture the complex contents of HRRS scenes. Also, an attention consistent model is developed to mitigate the influence caused by the spatial rotation. Apart from conventional DCNN, the attention mechanism also can be applied in generative adversarial networks (GANs). For instance, attention GANs [26] integrate the attention to further improve the representation power of the discriminator for enhancing the classification accuracy.

Apart from the attention mechanism, many other well-designed methods can be employed to assist the network to capture the semantic information within HRRS images. For example, the probabilistic topic model [48] was combined with CNNs to discover more discriminative semantics in [49], where deep features and hand-crafted features are fused adaptively to improve the results of scene classification. A multiple-instance densely-connected convnet (MIDC-Net) [50] was proposed to classify HRRS scenes. MIDC-Net transforms the HRRS image classification task into the multiple-instance learning problem, which helps the network explore target-level semantics. To further improve the classification performance, Zhang et al. [51] explored the semantic information distributions within the training and testing data, and they proposed a correlation subspace dynamic distribution alignment (CS-DDA) model to balance feature distributions between the source and target domains. Although the above HRRS scene classification methods achieve well performance, it is worth noting that the single scene label is hard to describe diverse semantics within HRRS images. Therefore, to explore the various semantics within HRRS images under the single-annotation scenario, a multi-granularity canonical appearance pooling (MG-CAP) [4] was proposed to capture the latent ontological structure of HRRS image data sets. MG-CAP progressively crops the input image to obtain multi-grained deep features, and a maxout-based Siamese style architecture is utilized to learn each deep feature with specific granularity. Finally, Gaussian covariance matrices and the matrix normalization method are employed to improve the discriminative power of features.

Due to the capacity of analyzing the unstructured data, GCN is becoming popular in the RS community. A growing number of GCN-based models have been developed for different RS tasks, such as hyperspectral image classification [52], [53] and HRRS scene classification [29], [30]. In [29], a deep feature aggregation framework driven by GCN was proposed. It utilizes a GCN to exploit patch-to-patch correlations of convolutional feature maps effectively. Thus, more refined features can be generated for HRRS image scene classification. Gao et al. [30] applied the high-order GCN to analyze the dependencies between different semantic classes. Different from the traditional GCN, high-order GCN can capture semantic dependencies at different orders. Therefore, the obtained feature representation is informative and discriminative, which improves classification performance.

Although the mentioned GCN-based methods have achieved positive classification results, two disadvantages limit their performance distinctly. First, due to the mechanism of graph theory, each pixel within RS images cannot be regarded as the graph node, resulting in substantial computational and storage costs. Second, to mitigate the above problems, researchers always select superpixels to be the graph nodes. However, this would lose the spatial context information, which is essential to the RS image interpretation. To overcome the above-mentioned limitations, our SAGN first constructs the graph at the superpixel level. Then, based on the similarities between different superpixels, we develop a dynamic updating approach to learn features from RS images at the pixel level. In this way, the time and storage costs can be reduced, and the spatial context information hidden in RS images can be preserved.

III. The Proposed Approach

This paper proposes an SAGN model to deal with HRRS scene classification tasks. The key idea of SAGN is to explore multiple semantics hidden in HRRS scenes. Furthermore, by mining their short-/long-range relationships the final decision of HRRS scenes can be obtained. The framework of SAGN is shown in Fig. 3, which consists of a dense feature pyramid network (DFPN), an adaptive semantic analysis module (ASAM), a dynamic graph feature update module (DGFUM), and a scene decision module (SDM) based on semantic regions. DFPN aims to extract the representative visual features from HRRS scenes, ASAM pays attention to analyze the latent and various semantics within HRRS scenes and generate diverse semantic regions adaptively, DGFUM updates the visual features under the paradigm of GCN to improve their discrimination with the consideration of the relationships between semantic regions, and SDM utilizes the updated features and corresponding semantic regions to decide the final semantic labels of HRRS scenes.

A. Dense Feature Pyramid Network

Feature extraction is a challenging task due to the complex contents of HRRS scenes. To obtain the representative features, DFPN is developed in this section, which adopts ResNet [54] as backbone to generate multi-scale features. The flowchart of DFPN is shown in Fig. 4.

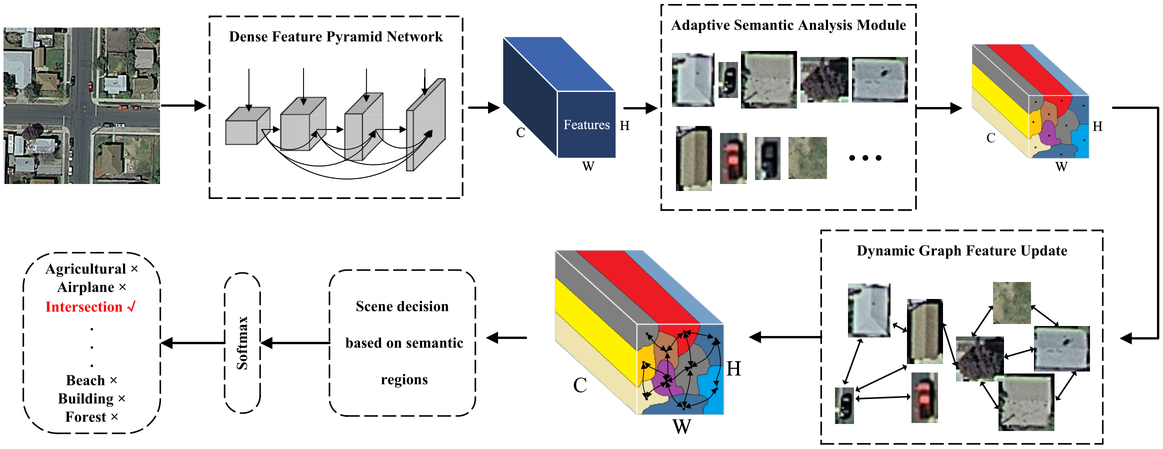

Dense Feature Pyramid Network
Features
C
H
W
Adaptive Semantic Analysis Module
...
Dynamic Graph Feature Update
Scene decision based on semantic regions
Softmax
Agricultural ×
Airplane ×
Intersection √
Beach ×
Building ×
Forest ×

Fig. 3. The framework of SAGN, which consists of four parts: a dense feature pyramid network, an adaptive semantic analysis module, a dynamic graph feature update module, and a scene decision module based on semantic regions.

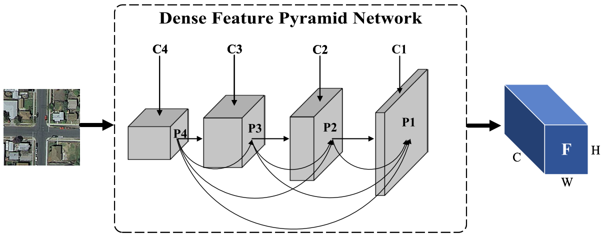


Fig. 4. Flowchart of dense feature pyramid network.

When an HRRS image is input to DFPN, four feature maps C_1, C_2, C_3, and C_4 that contain multi-scale semantic information are generated by four residual blocks of ResNet. Then, the bottom-up structure is applied to fuse them. We can formulate this process as

$$\begin{aligned} P_4 &= Conv_{1\times1}(C_4), \\ P_i &= Conv_{3\times3}[\sum_{j=i+1}^{4} Upsample(P_j) \oplus Conv_{1\times1}(C_i)], \end{aligned} \tag{1}$$

where $Conv_{k\times k}(\cdot)$ denotes the convolution operation with a kernel size of $k \times k$, $Upsample(\cdot)$ represents the bilinear upsampling, and $\oplus$ indicates the concatenation operation. Finally, P_1 is adopted as the final visual feature $\mathbf{F}$.

B. Adaptive Semantic Analysis Module

Diverse semantics always appear in an HRRS scene. The pre-defined single semantic label can represent the main content in this scene. However, other semantics (which are ignored by the manual label) can also provide useful information for distinguishing the scene. Analyzing the diverse semantics within an HRRS scene and mining the context relations between them are beneficial to the classification task. Therefore, we propose ASAM to capture latent semantics and generate diverse semantic regions adaptively such that the various information can be explored for deciding the label of the HRRS scene.

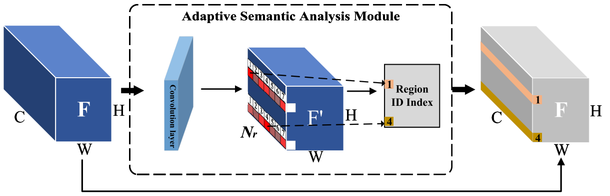


Fig. 5. Schematic illustration of the adaptive semantic analysis module. The diverse semantics within an HRRS scene can be explored and represented by different semantic regions. Here, only two feature points are analyzed and exhibited for clear.

Here, the number of semantic regions N_r needs to be set in advance for ASAM. In this paper, its value is up to the class number of the data set. To capture the latent semantics of HRRS image efficiently, ASAM adopts a simple structure that contains one convolution layer with the kernel size of 5×5. Note that the outputting channel number of convolution layer equals N_r. The schematic illustration of semantics analysis is shown in Fig. 5. Specifically, $\mathbf{F}$ is first convoluted to $\mathbf{F}'$ by the first convolution layer with the kernel size of 5×5. Here, the channel number of $\mathbf{F}'$ is equal to N_r. Second, for each feature point of $\mathbf{F}'$, the channel-wise position with the maximal value is recorded as the ID of semantic regions. Finally, the semantic regions of $\mathbf{F}$ can be obtained according to each point's ID index adaptively.

C. Dynamic Graph Feature Update Module

Before introducing our DGFUM, the preliminaries of GCN will be discussed. GCN is a popular tool for processing unstructured data, which aims to extract advanced features by aggregating the information from graph nodes' neighborhoods. Suppose there is a set of graph nodes $\mathbf{X}_g = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{n_g}\}$, where $\mathbf{x}_i$ denotes the i-th graph node, and n_g equals the number of graph nodes. To describe the interrelationship between graph nodes (i.e., the edges between graph nodes), the adjacency matrix $\mathbf{A}$ is defined as,

$$\mathbf{A}_{ij} = \begin{cases} e^{-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2}, & \text{if } \mathbf{x}_i \in N(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in N(\mathbf{x}_i) \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

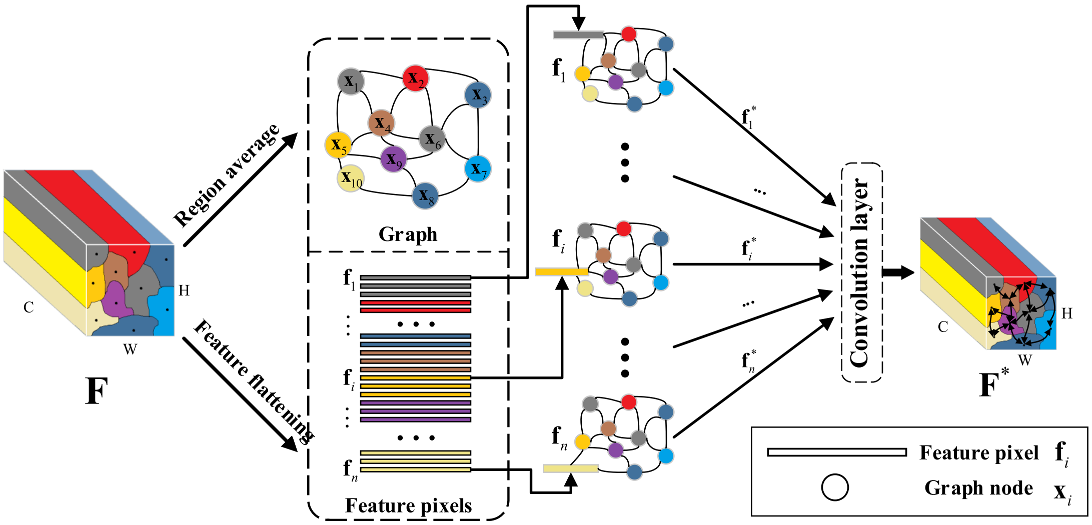


Fig. 6. Schematic illustration of the dynamic graph feature update module. In this example, visual feature map $\mathbf{F}$ contains n feature pixels and is analyzed into ten semantics, where n equals $W \times H$. Therefore, feature $\mathbf{F}$ can be regionally averaged to ten graph nodes $\mathbf{x}_1$ to $\mathbf{x}_{10}$ and flatted to n feature pixels $\mathbf{f}_1$ to $\mathbf{f}_n$. The edges of graph (i.e., the regional adjacency matrix) can be generated by measuring the similarity relationships between different nodes. Then, each feature pixel $\mathbf{f}$ can be updated to $\mathbf{f}^*$ by the graph convolution with the proposed graph node updating strategy with the region-level adjacency relations. Finally, after a convolution layer, all the updated feature pixels $\mathbf{f}^*$ compose the updated visual feature $\mathbf{F}^*$.

where γ is a hyper-parameter, $\mathbf{x}_i$ and $\mathbf{x}_j$ denote two graph nodes, $N(\mathbf{x}_i)$ indicates the neighbor set of $\mathbf{x}_i$. To update the nodes' representation, the learnable weight $\mathbf{W}$ is introduced, and the process of graph convolution of the l-th layer can be defined as,

$$\mathbf{X}_g^{(l)} = \sigma(\mathbf{A}^{(l)} \cdot \mathbf{X}_g^{(l-1)} \cdot \mathbf{W}^{(l)}), \tag{3}$$

where $\mathbf{X}_g^{(l)}$ and $\mathbf{X}_g^{(l-1)}$ are the output and input of the l-th layer, $\sigma(\cdot)$ denotes the activation function, and $\mathbf{A}^{(l)}$ and $\mathbf{W}^{(l)}$ denote the adjacency matrix and trainable matrix of the l-th layer, respectively.

Now, let us explain DGFUM in detail. Under the GCN framework, our DGFUM is developed for the specific scenario, i.e., HRRS scene understanding. As mentioned in Section III-B, an HRRS scene has been analyzed into several semantic regions. Thus, we regard those regions as the nodes to construct a graph. Specifically, suppose the visual features $\mathbf{F} \in \mathbb{R}^{W\times H\times C}$ have been divided into N_r semantic regions adaptively, where $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, \ldots, \mathbf{f}_n\}$, $\mathbf{f}_i \in \mathbb{R}^C$ and $n = W \times H$. We use the center feature to represent each region, which can be calculated by averaging the features (in the channel dimension) within a region. Then, a set of graph nodes $\mathbf{X}_g = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_{N_r}\}$ can be obtained, where $\mathbf{x}_i \in \mathbb{R}^C$ and C is the channel number of the visual features $\mathbf{F}$. To deeply capture their interrelationship, a trainable adjacency matrix $\mathbf{A}$ is adopted, whose definition is

$$\mathbf{A}_{ij} = e^{-\gamma\cdot(\mathcal{D}(\mathbf{x}_i,\mathbf{x}_j))^2}, \tag{4}$$

where γ is a hyper-parameter, $\mathbf{x}_i$ and $\mathbf{x}_j$ denote two graph node features, $\mathcal{D}(\cdot)$ means the distance metric. In this paper, instead of using the Euclidean distance metric, we adopt the Mahalanobis distance metric to measure the resemblance between nodes which can be learned during the graph convolutional process. Its definition is

$$\mathcal{D}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^\top \mathbf{M}(\mathbf{x}_i - \mathbf{x}_j)}, \tag{5}$$

where $\mathbf{M}$ can be decomposed as $\mathbf{W}_d\mathbf{W}_d^\top$ and $\mathbf{W}_d$ is a trainable weight matrix.

Next, different from the conventional GCN model that focuses on updating nodes' features, our DGFUM aims to update the feature points within $\mathbf{F}$ with the new graph node updating strategy. In other words, DGFUM aims at transforming $\mathbf{F}$ to $\mathbf{F}^*$ through the graph convolution with the region-level adjacency relations decided by $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{N_r}\}$. In detail, for one feature point $\mathbf{f}_i$ of $\mathbf{F}$, if it belongs to the j-th semantic region, its renewing process can be formulated as,

$$\mathbf{f}_i^* = \sigma(\mathbf{A}_j \cdot \begin{bmatrix} 1: & \mathbf{x}_1 \\ 2: & \mathbf{x}_2 \\ 3: & \mathbf{x}_3 \\ & \cdots \\ j: & \mathbf{f}_i \\ & \cdots \\ n: & \mathbf{x}_{N_r} \end{bmatrix}), \tag{6}$$

where $\mathbf{f}_i$ and $\mathbf{f}_i^*$ are the original and updated features, $\sigma(\cdot)$ denotes the activation function, $\mathbf{A}_j$ represents the j-th row of adjacency matrix $\mathbf{A}$, $\mathbf{x}_i$ means the i-th graph node, and $j : \cdot$ denotes the j-th row of matrix. After all the feature points are updated by this update strategy, a convolution layer with the kernel size of 3×3 is applied to the updated $\mathbf{F}$ for the final feature $\mathbf{F}^*$. The schematic illustration of update process is shown in Fig. 6. The reason why we add a convolution operation for updated features is inspired by the second matrix multiplication of traditional GCN (i.e., $\mathbf{F}_g^{(l-1)} \cdot \mathbf{W}^{(l)}$ of Eq. 3). This operation of traditional GCN is the same as the convolution operation with the kernel size of 1×1, which can capture the relationship between the different channels. Furthermore, in addition to the channel relationship, the convolution layer with the kernel size of 3 × 3 can also capture the local information, which can complement the global information of GCN.

On the one hand, in comparison with CNN, DGFUM can capture global and local information within HRRS scenes by utilizing GCN to fuse all the region features with different semantics. This could increase the discrimination of learned features. On the other hand, rather than update the regional feature representation $\mathbf{F}_g$, we learn the point features $\mathbf{F}$ with the assistance of the regional adjacency matrix. To sum up, DGFUM can learn the feature point approximately with the low computational costs. Therefore, the semantic labels of HRRS scenes can be predicted rapidly and accurately.

D. Scene Decision Based on Semantic Regions

The updated visual features $\mathbf{F}^*$ contain discriminative information so that HRRS scenes can be distinguished accurately. To obtain the semantic labels of HRRS scenes rapidly and precisely, a convolution layer with the kernel of 1 × 1 is adopted to process $\mathbf{F}^*$. The output channel equals the number of semantic classes in the HRRS scene data set. Then, the softmax function is added to normalize the output in the channel dimension. Finally, the scene class corresponding to the channel-wise position with the maximal value is selected as the final label of an HRRS scene. It is worth noting that only the cross-entropy is employed as the loss function to train the network in the training stage.

IV. Experimental Results

A. Data Set Introduction

Three public data sets are employed to evaluate our SAGN, including UC Merced (UCM) data set [25], AID data set [55], and NWPU-RESISC45 data set [56].

1) UC Merced Data Set: UCM contains 2100 HRRS images with diverse land-use patterns which are obtained from United States Geological Survey National Map of several U.S. regions, and they are divided into 21 semantic classes equally. The size and spatial resolution are 256 × 256 × 3 and 0.3m. The examples are shown in Fig. 7, which are selected from each class randomly.

2) AID Data Set: AID contains 10000 HRRS images which are collected from the Google Earth Imagery. There are 30 scene classes and the sample number of each class varies from 220 to 420. The size and spatial resolution are 600 × 600 and from 0.5m to approximate 8m. The image examples of data set are shown in Fig. 8.

3) NWPU-RESISC45 Data Set: There are 31500 HRRS images in this data set which are divided into 45 scene classes equally. The size and spatial resolution of each HRRS image are 256 × 256 and 30-0.2m per pixel, respectively. The examples are shown in Fig. 9.

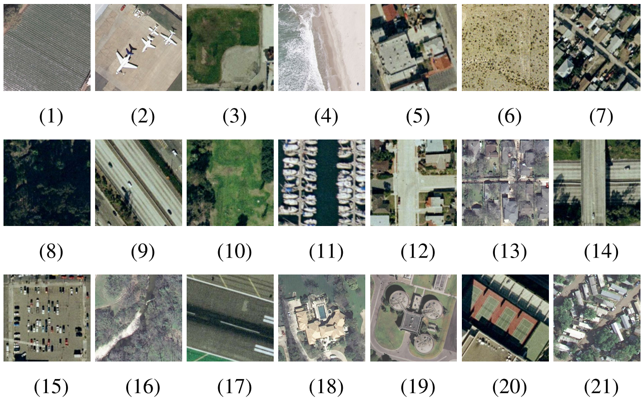


Fig. 7. Examples in the UCM data set: (1) Agricultural ∼ 100, (2) Airplane ∼ 100, (3) Baseball Diamond ∼ 100, (4) Beach ∼ 100, (5) Buildings ∼ 100, (6) Chaparral ∼ 100, (7) Dense Residential ∼ 100, (8) Forest ∼ 100, (9) Freeway ∼ 100, (10) Golfcourse ∼ 100, (11) Harbor ∼ 100, (12) Intersection ∼ 100, (13) Medium Residential ∼ 100, (14) Overpass ∼ 100, (15) Parking Lot ∼ 100, (16) River ∼ 100, (17) Runway ∼ 100, (18) Sparse Residential ∼ 100, (19) Storage Tanks ∼ 100, (20) Tennis Court ∼ 100, (21) Mobile Home Park ∼ 100. (Semantic category ∼ Number of images.)

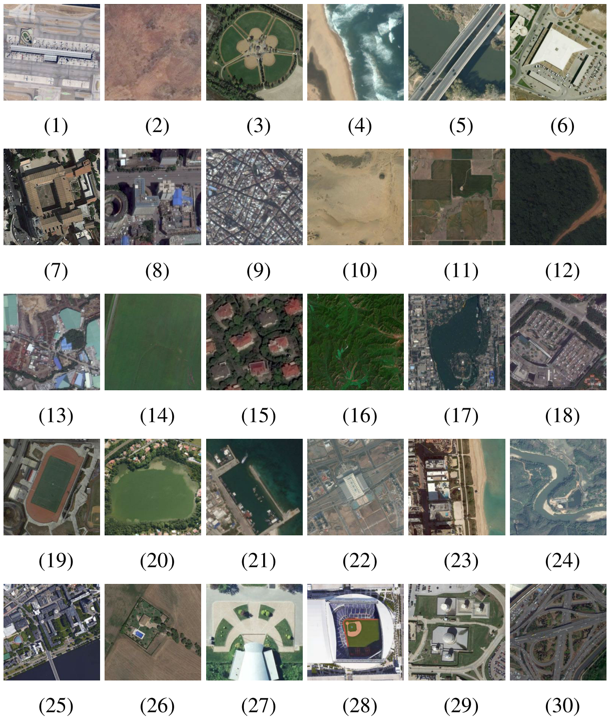


Fig. 8. Examples in the AID data set: (1) Airport ∼ 360, (2) Bare Land ∼ 310, (3) Baseball Field ∼ 220, (4) Beach ∼ 400, (5) Bridge ∼ 360, (6) Center ∼ 260, (7) Church ∼ 240, (8) Commercial ∼ 350, (9) Dense Residential ∼ 410, (10) Desert ∼ 300, (11) Farmland ∼ 370, (12) Forest ∼ 250, (13) Industrial ∼ 390, (14) Meadow ∼ 280, (15) Medium Residential ∼ 290, (16) Mountain ∼ 340, (17) Park ∼ 350, (18) Parking ∼ 390, (19) Playground ∼ 370, (20) Pond ∼ 420, (21) Port ∼ 380, (22) Railway Station ∼ 260, (23) Resort ∼ 290, (24) River ∼ 410, (25) School ∼ 300, (26) Sparse Residential ∼ 300, (27) Square ∼ 330, (28) Stadium ∼ 290, (29) Storage Tanks ∼ 360, (30) Viaduct ∼ 420. (Semantic category ∼ Number of images.)

B. Implementation Details

The training and inference are conducted by pytorch [57] on a high performance computer with GeForce RTX 3090 of 24G memory and Inter Xenon Silver 4214R. In the training stage, the ResNet of DFPN are initialized by the pre-trained parameters (using ImageNet data set [36]) and the rest parts of SAGN are initialized by a set of random parameters which follow a normal distribution with a standard deviation of 0.1. To train SAGN, we employ the Adam optimizer

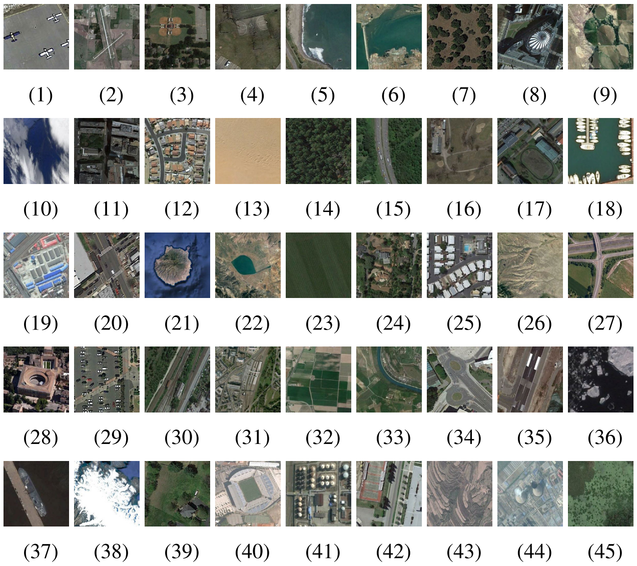


Fig. 9. Examples in the NWPU data set: (1) Airplane ∼ 700, (2) Airport ∼ 700, (3) Baseball Diamond ∼ 700, (4) Basketball Court ∼ 700, (5) Beach ∼ 700, (6) Bridge ∼ 700, (7) Chaparral ∼ 700, (8) Church ∼ 700, (9) Circular Farmland ∼ 700, (10) Cloud ∼ 700, (11) Commercial Area ∼ 700, (12) Dense Residential ∼ 700, (13) Desert ∼ 700, (14) Forest ∼ 700, (15) Freeway ∼ 700, (16) Golf Course ∼ 700, (17) Ground Track Field ∼ 700, (18) Harbor ∼ 700, (19) Industrial Area ∼ 700, (20) Intersection ∼ 700, (21) Island ∼ 700, (22) Lake ∼ 700, (23) Meadow ∼ 700, (24) Medium Residential ∼ 700, (25) Mobile Home Park ∼ 700, (26) Mountain ∼ 700, (27) Overpass ∼ 700, (28) Palace ∼ 700, (29) Parking lot ∼ 700, (30) Railway ∼ 700, (31) Railway Station ∼ 700, (32) Rectangular Farmland ∼ 700, (33) River ∼ 700, (34) Roundabout ∼ 700, (35) Runway ∼ 700, (36) Sea Ice ∼ 700, (37) Ship ∼ 700, (38) Snow Berg ∼ 700, (39) Sparse Residential ∼ 700, (40) Stadium ∼ 700, (41) Storage Tank ∼ 700, (42) Tennis Court ∼ 700, (43) Terrace ∼ 700, (44) Thermal Power Station ∼ 700, (45) Wetland ∼ 700. (Semantic category ∼ Number of images.)

with 32 images per minibatch and train the network with 1000 epochs. The initial learning rate of network is 4×10^{-5}, and it is divided by 10 after each 200 epochs.

In addition, the data augmentation is employed to enhance the robustness of the model. First, the short side length of the HRRS image is resized to 256 while the ratio of width and height is kept. Second, the resized image is randomly rotated within the angle range of $[-5, +5]$. Third, the rotated image would be randomly horizontally flipped with a probability of 0.5. Fourth, we randomly crop 256 × 256 patches from the resized image. Finally, the cropped image would be normalized by subtracting mean and dividing standard deviation. Similar to the papers [4], [29], [30], [58], the ratios of training set of UCM, AID and NWPU data sets are 80%, 20% and 50%, and 10% and 20%, respectively.

C. Evaluation Metrics

To evaluate the performance of our SAGN, four assessment criteria are utilized, including the overall accuracy (OA) [29], average accuracy (AA) [29], Kappa coefficient (Kappa) [29], and the confusion matrix (CM) [59]. Here, OA is defined as the number of correctly classified images divided by the number of total testing images. AA is the average of the precision of all classes. Kappa is defined to measure the consistency between the prediction results and ground truth. CM is a detailed table in which the column indicates the ground truth and the row denotes the prediction. According to the CM, we can easily find if the predicted labels of the test data are correct or not.

D. Analysis of SAGN

1) The Influence of N_r to SAGN: As mentioned in Section III-B, there is a hyper-parameter N_r in SAGN, which controls the number of semantic regions for HRRS images. To study its influence on our model, we change its values and review the performance of SAGNs. In particular, we vary N_r from 1 to the number of classes in the data set with the approximate exponential interval. Taking the UCM data set as an example, we set $N_r = 1, 8, 16, 21$, respectively. Here, other experimental settings are the same as the contents mentioned in Section IV-B. The results of different SAGNs counted on three data sets are summarized in Table I, and the optimal OA values are marked in bold. Note that if the highest OA value corresponding to different N_r values, the minimum N_r value would be selected as the optimal N_r with the consideration of computational costs.

From the observation of the results, the performance of our model is varied within an acceptable range based on different N_r, and the optimal value of N_r is directly proportional to the complexity of the data set. In particular, for UCM, when N_r=8, our model achieves the highest OA value (i.e., 99.76%). Then, with the increase of N_r, the OA values are kept. Therefore, the optimal N_r value is equal to 8. For the AID data set, when N_r=16, SAGNs perform the best in all cases we have tried thus far. It is worth noting that OA values decrease when N_r increases from 16 to 30, which indicates that the behavior of SAGNs does not get better with the larger N_r. For the NWPU data set, like AID, when N_r=16, SAGNs reach the highest accuracy in two scenarios, i.e., 91.75% and 93.53%. Accordingly, the optimal value of semantic region number N_r is set at 8, 16, and 16 for UCM, AID, and NWPU in the following experiments unless otherwise specified.

We want to touch on another point, i.e., when the number of semantic regions N_r equals 1, ASAM and DGFUM would be invalid. In other words, when N_r increases, the contributions of the two mentioned modules can be explored. From the above discussion, we can easily find that all optimal OA values (for three data sets) are obtained when $N_r > 1$. It means the usefulness of ASAM and DGFUM to HRRS scene classification tasks modules distinctly.

2) Performance of SAGN: To study the performance of SAGN, we count its OA, AA, and Kappa values on three data sets. The results are exhibited in Table II. For UCM, we find that the values of OA and AA are the same, and Kappa is high (0.9975). This is because only one image is mispredicted. It demonstrates that SAGN performs well for the UCM data set. For AID, when the training ratio equals 20%, OA, AA, and Kappa of SAGN equal 95.35%, 94.98%, and 0.9507, respectively. This is because the numbers of RS images corresponding to different categories are different. Furthermore, the distinct numerical gap between OA and AA indicates that the behavior of our model is not strong for each category. When the training ratio is equal to 50%, the

TABLE I
THE OA CORRESPONDING TO DIFFERENT N_r VALUE ON UCM, AID, NWPU DATA SETS (%)

UCM		AID				NWPU			
T.R. = 80%		T.R. = 20%		T.R. = 50%		T.R. = 10%		T.R. = 20%	
N_r	OA	N_r	OA	N_r	OA	N_r	OA	N_r	OA
1*	99.52	1*	94.92	1*	96.32	1*	91.37	1*	93.09
8	**99.76**	8	95.21	8	96.66	8	91.55	8	93.37
16	99.76	16	**95.35**	16	**96.90**	16	**91.75**	16	**93.53**
21	99.76	30	95.01	30	96.86	32	91.66	32	93.46
-	-	-	-	-	-	45	91.70	45	93.53

TABLE II
THE EVALUATION FOR UCM, AID, NWPU DATA SETS (%)

UCM		AID				NWPU			
T.R. = 80%		T.R. = 20%		T.R. = 50%		T.R. = 10%		T.R. = 20%	
OA	99.76	OA	95.35	OA	96.90	OA	91.75	OA	93.53
AA	99.76	AA	94.98	AA	96.70	AA	91.75	AA	93.53
Kappa	0.9975	Kappa	0.9507	Kappa	0.9679	Kappa	0.9160	Kappa	0.9338

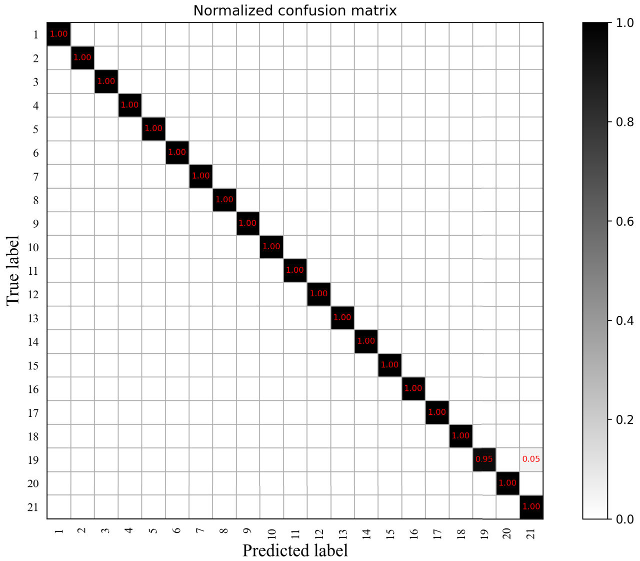


Fig. 10. The confusion matrix of SAGN on UCM data set when the train ratio equals 80% and OA equals 99.76%. The semantic names corresponding to different numbers can be found in Fig. 7.

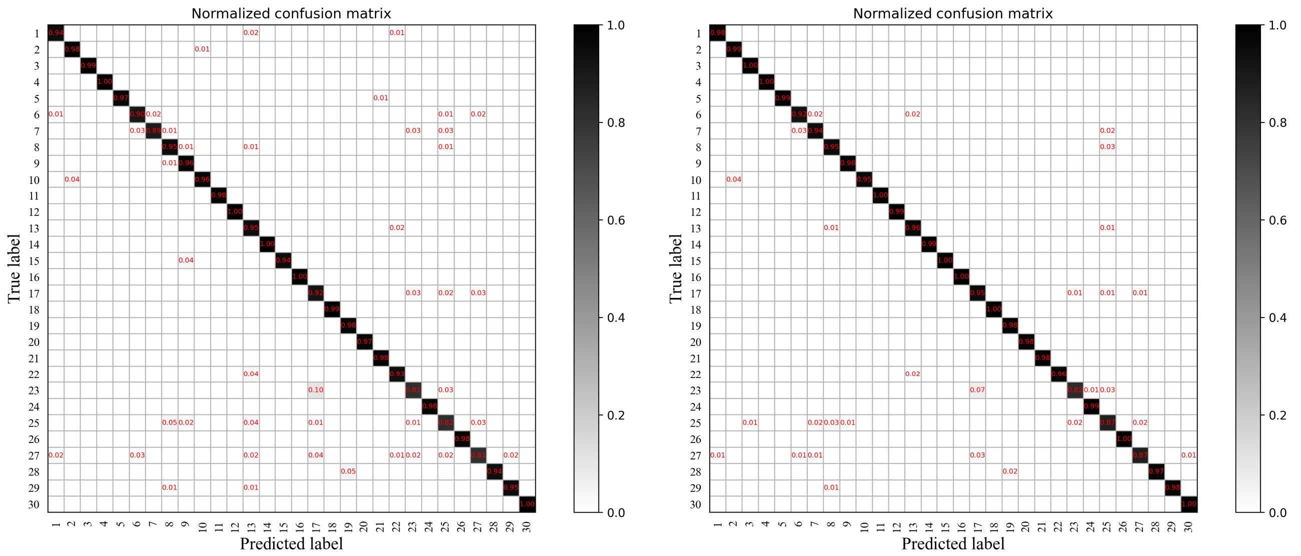
gap between assessment criteria is reduced. This implies that more training data can improve the SAGN's performance for different categories. For NWPU, when the training ratio equals 10% and 20%, our model's OA and AA values are the same. It confirms that SAGN has stable performance in each class of the NWPU data set, which is also demonstrated by the high Kappa.

To study the behavior of SAGN for different semantic categories, we count its CMs on the UCM, AID, and NWPU data sets, which are displayed in Figs. 10, 11, and 12, respectively. For UCM, according to Fig. 10, it is easily find that only 5% HRRS scenes which belong to "Storage Tanks" are classified to "Moblie Home Park" incorrectly. For the rest scenes, SAGN has the stable and accurate performance.

For AID, the CM of SAGN counted by the results with different proportions of training samples are shown in Fig. 11. According to the figures, when there are 20% training samples, SAGN reaches the 100% (in OA) on the following five scenes, including "Beach", "Forest", "Meadow", "Mountain", and "Viaduct". This confirms the usefulness of our method. However, its OA values on "Resort" and "Square" are relatively low, which are only 81%. For the "Resort" scene, SAGN groups 10% samples into "Park" incorrectly because "Resort" and "Park" scenes have similar semantic distributions, which would mislead our model. For the "Square" scene, its diverse semantic contents disturb SAGN so that some samples are divided into various scenes, such as "Airport", "Center", "Industrial", etc. When the ratio of training set increases to 50%, SAGN can achieve 100% (in OA) on more scenes, including "Beach", "Farmland", "Medium Residential", "Mountain", "Parking", "Sparse Residential", and "Viaduct". However, the performance of SAGN on the "Resort" scene is still not as good as we expect due to the high semantic similarity between "Resort" and "Park" scenes.

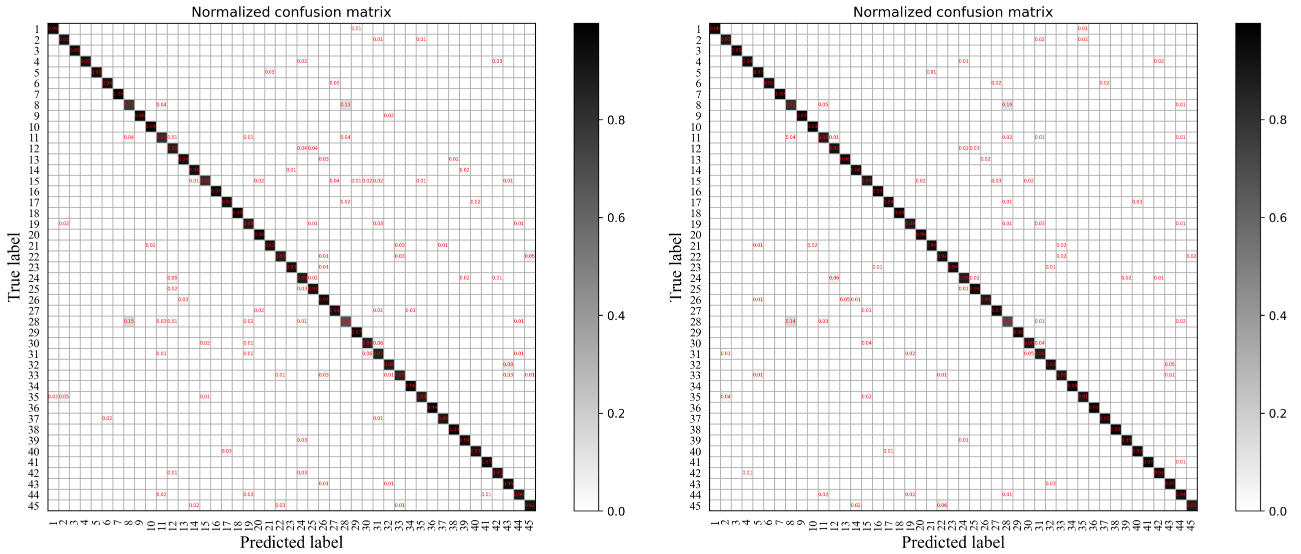
For NWPU, CM corresponding to different training sets are counted and shown in Fig. 12. According to Fig. 12a, when 10% data set are utilized to be training set, SAGN achieves the highest accuracy 100% on "Chaparral" and "Snow Berg" scenes. Similarly, when the training set ratio is set to be 20% (see Fig. 12b), SAGN obtains the best accuracy (100%) on "Chaparral" scene. It benefits from the semantic mining and short-/long-range context exploration of our model. However, SAGN is not good at distinguishing the "Palace" scene, where the OA values are 68% and 71% in two cases, and the most incorrect predicted samples appear in "Church". It is because that the contents within these two scenes are similar in type and distribution, which prevents our model from dividing them successfully.

3) Analysis of ASAM: As discussed in Section III, an RS scene always contains complex contents that include semantics corresponding to the pre-defined label and various types of land-covers. Considering them and their context information simultaneously would be beneficial to feature learning and classification. To this end, we propose ASAM to discover the possible semantics hidden in an RS scene. ASAM is a simple model, which only contains a convolution and channel-



(a) (b)

Fig. 11. The confusion matrix of SAGN on AID data set. The semantic names corresponding to different numbers can be found in Fig. 8. (a) The ratio of training set equals 20% and OA is equal to 95.21%. (b) The ratio of training set equals 50% and OA is equal to 96.90%.



(a) (b)

Fig. 12. The confusion matrix of SAGN on NWPU data set. The semantic names corresponding to different numbers can be found in Fig. 9. (a) The ratio of training set equals 10% and OA is equal to 91.75%. (b) The ratio of training set equals 20% and OA is equal to 93.53%.

wise position selection operations (see Fig. 5). To study its usefulness, we do the following experiments.

In the first place, we visualize the result of ASAM to study if the possible semantics can be explored or not. Specifically, we randomly select a "Dense Residential" RS scene from the UCM data set (Fig. 13a). Then, we pass it through a simple network (i.e., three convolution layers) and ASAM to get the semantic exploration result, as shown in Fig. 13b. At the same time, the k-means result based on the features obtained after 5 × 5 convolution is also exhibited for reference. It can be seen that: (1) ASAM can capture various semantics within the RS scene. In Fig. 13b, apart from "Dense Residential" (cyan), many other semantics have been captured, such as "Freeway" (blue), "Forest" (green), and "Buildings" (red). (2) The obtained semantic regions are similar to the results obtained by k-means (Fig. 13c), which confirms the reliability of our semantic exploration result. Not that, the reason why we do not embed the k-means algorithm here is our ASAM should be as simple as possible. In other words, ASAM should be efficient because it is only a part of SAGN.

In the second place, we study the influence of different kernels on ASAM. Particularly, we select 1 × 1 and 5 × 5 convolution kernels to complete ASAM, and the results are exhibited in Fig. 14. From the observation of them, we can find that the semantic regions of ASAM with 5×5 convolution (Fig. 14a) are more clear than that of ASAM with 1 × 1

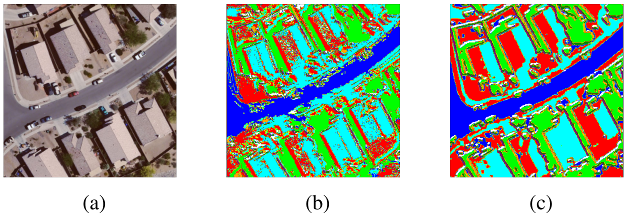


Fig. 13. Visual results of semantic analysis obtained by ASAM. (a) A "Dense Residential" RS scene from the UCM data set. (b) The analyzed result generated by ASAM. (c) The clustering result generated by k-means.

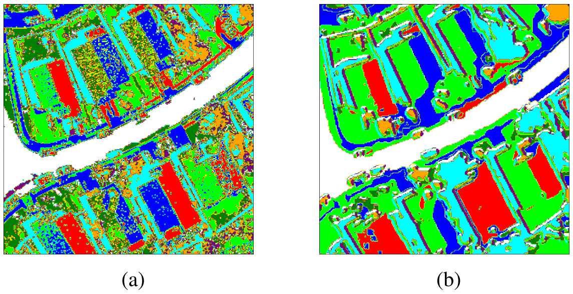


Fig. 14. Visual results of semantic analysis obtained by ASAM with different convolutional kernels. (a) The kernel size is 1 × 1. (b) The kernel size is 5 ×5.

(Fig. 14b). Besides the uniform local information, smooth edges can also be discovered. This is because convolutional kernels with a size of 5 × 5 are aware of the spatial neighborhood relationships. Thus, in our ASAM, we set the convolutional kernel size at 5 × 5.

Last but not least, we study the influence of N_r on the model consisting of DFPN and ASAM visually. In particular, we select two RS scenes from the UCM data set randomly, which are from "River" and "Parking Lot" categories (Fig. 15a). Then, we set N_r at 8, 16, and 21 to get ASAM results, respectively, which are shown in Figs. 15b, 15c, and 15d. According to the results, we can find that when N_r is varied, the number of possible semantics is increased. Let us take "River" as an example. When N_r equals 8, "River," "Forest," and other semantic regions can be discovered (see Fig. 15b). However, when N_r increases, the obtained regions are fragmented, and their discriminative information is reduced. Thus, the value of N_r should be tuned carefully. How to decide an appropriate value of N_r adaptively is one of our future works.

4) Study of Distance Metric: As mentioned in Section III-C, since the Mahalanobis distance metric can be transformed into the learnable version (see Eq. 5), it is selected to measure the resemblance between graph nodes in our DGFUM. In this section, we study its usefulness. In particular, we construct two SAGNs. One is embedded with the Mahalanobis distance metric, and the other is embedded with the Euclidean distance metric. They are recorded SAGN-M and SAGN-E, respectively. Then, two models are trained and testified by three data sets. Here, we randomly choose 80%, 50%, and 20% RS images from UCM, AID, and NWPU as the training data. The OAs of two SAGNs counted on different data sets are shown in Table III. After studying the results, SAGN embedded by the learnable Mahalanobis distance achieves the stronger behavior. It illustrates the superiority of the selected learnable distance metric.

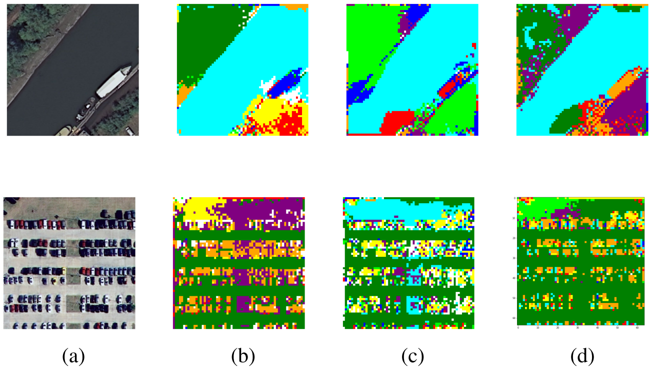


Fig. 15. Visual results of semantic analysis obtained by ASAM with different N_r. (a) The original images (Top: "River", Bottom: "Parking Lot"). (b) Analysis result with N_r equals 8. (c) Analysis result with N_r equals 16. (d) Analysis result with N_r equals 21 (the class number of UCM data set).

TABLE III

OVERALL ACCURACY OF SAGNS WITH THE LEARNABLE MAHALANOBIS DISTANCE AND THE EUCLIDEAN DISTANCE. ALL RESULTS ARE COUNTED ON THE UCM, AID, NWPU DATA SETS (%)

	UCM (T.R. = 80%)	AID (T.R. = 50%)	NWPU (T.R. = 20%)
SAGN-M	**99.76**	**96.90**	**93.53**
SAGN-E	99.57	96.69	93.41

TABLE IV

OVERALL ACCURACY OF SAGNS WITH THE PROPOSED AND TRADITIONAL GRAPH NODE UPDATING STRATEGIES ON THE UCM, AID, NWPU DATA SETS (%)

	UCM (T.R. = 80%)	AID (T.R. = 50%)	NWPU (T.R. = 20%)
SAGN-M	**99.76**	**96.90**	**93.53**
SAGN-E	97.74	95.07	92.63

5) Analysis of the Proposed Graph Node Updating Strategy: In this section, we study if the proposed graph node updating strategy is good for our scene classification tasks or not. Particularly, two SAGNs are constructed with the proposed and traditional graph node updating strategies to complete the HRRS scene classification. We use SAGN-P to SAGN-T to denote them for short. Their results on three data sets are summarized in Table IV, in which the 80%, 50%, and 20% samples are chosen randomly to train two SAGNs. From observing the results, we find that SAGN-P outperforms SAGN-T in all cases. The improvements obtained by SAGN-P compared to SAGN-T are 2.02%, 1.83%, and 0.9%, respectively. The positive results imply that the proposed graph node updating strategy is helpful to our tasks.

E. Comparison With Some State-of-the-Art Methods

To study the performance of SAGN, fifteen models (eight deep-based methods, five their variants, and two graph-based methods) are employed, including AlexNet+SVM [59], MSCP$_1$ [58], MSCP+MRA$_1$ [58], DCNN$_1$ [59], GoogleNet+SVM [59], DCNN$_2$ [59], MSCP$_2$ [58], VGG-D+SVM [59], MSCP+MRA$_2$ [58], RTN [28], MG-CAP with Sqrt-E [4], Context aggregation [40], Binary pattern

TABLE V

THE FEATURE EXTRACTOR OF ALL METHODS IN OUR EXPERIMENT. (*: CNN-BASED METHOD; †: GCN-BASED METHOD)

Methods	Feature Extractor
AlexNet+SVM [59] *	AlexNet
MSCP$_1$ [58] *	AlexNet
MSCP+MRA$_1$ [58] *	AlexNet
DCNN$_1$ [59] *	AlexNet
GoogleNet+SVM [59] *	GoogLeNet
DCNN$_2$ [59] *	GoogLeNet
MSCP$_2$ [58] *	VGG-D
VGG-D+SVM [59] *	VGG-D
MSCP+MRA$_2$ [58] *	VGG-D
RTN [28] *	VGG-D
MG-CAP with Sqrt-E [4] *	VGG-D
Context aggregation [40] *	ResNet
Binary pattern encoded CNNs [44] *	ResNet
DFGCN [29] †	VGG-D
H-GCN [30] †	DenseNet
SAGN (our) †	ResNet

TABLE VI

OVERALL ACCURACY AND STANDARD DEVIATIONS OF THE PROPOSED SAGN AND THE COMPARED METHODS ON THE UCM DATA SET (%). THE ENTRY WITH THE HIGHEST VALUE IS SHOWN IN BOLD. (*: CNN-BASED METHOD; †: GCN-BASED METHOD)

Methods	OA (T.R. = 80%)
AlexNet+SVM [59] *	94.42 ± 0.10
DCNN$_1$ [59] *	96.67 ± 0.10
GoogleNet+SVM [59] *	96.82 ± 0.20
DCNN$_2$ [59] *	97.07 ± 0.12
VGG-D+SVM [59] *	97.14 ± 0.10
MSCP$_1$ [58] *	97.29 ± 0.63
MSCP+MRA$_1$ [58] *	97.32 ± 0.52
Binary pattern encoded CNNs [44] *	97.72 ± 0.54
MSCP$_2$ [58] *	98.36 ± 0.58
MSCP+MRA$_2$ [58] *	98.40 ± 0.34
Context aggregation [40] *	98.56 ± 0.53
RTN [28] *	98.96
MG-CAP with Sqrt-E [4] *	99.00 ± 0.10
DFGCN [29] †	98.48 ± 0.42
H-GCN [30] †	99.00 ± 0.43
SAGN (our) †	**99.82** ± **0.10**

TABLE VII

OVERALL ACCURACY AND STANDARD DEVIATIONS OF THE PROPOSED SAGN AND THE COMPARED METHODS ON THE AID DATA SET (%). THE ENTRY WITH THE HIGHEST VALUE IS SHOWN IN BOLD. (*: CNN-BASED METHOD; †: GCN-BASED METHOD)

Method	OA	
	T.R. = 20%	T.R. = 50%
AlexNet+SVM [59] *	84.23 ± 0.10	93.51 ± 0.10
DCNN$_1$ [59] *	85.62 ± 0.10	94.47 ± 0.10
GoogleNet+SVM [59] *	87.51 ± 0.11	95.27 ± 0.10
DCNN$_2$ [59] *	88.79 ± 0.10	96.22 ± 0.10
MSCP$_1$ [58] *	88.99 ± 0.38	92.36 ± 0.21
VGG-D+SVM [59] *	89.33 ± 0.23	96.04 ± 0.13
MSCP+MRA$_1$ [58] *	90.65 ± 0.19	94.11 ± 0.15
MSCP$_2$ [58] *	91.52 ± 0.21	94.42 ± 0.17
MSCP+MRA$_2$ [58] *	92.21 ± 0.17	96.56 ± 0.18
MG-CAP with Sqrt-E [4] *	93.34 ± 0.18	96.12 ± 0.12
Binary partern encoded CNNs [44] *	93.81 ± 0.12	95.73 ± 0.16
DFGCN [29] †	90.47 ± 0.11	94.88 ± 0.22
H-GCN [30] †	93.06 ± 0.26	95.78 ± 0.37
SAGN (our) †	**95.17** ± **0.12**	**96.77** ± **0.18**

encoded CNNs [44], DFGCN [29], and H-GCN [30]. The feature extractor of these models are different. We summarize them and corresponding base type in Table V.

1) Results on UCM: The OA values of all models are summarized in Table VI. It is obvious that our model SAGN has the best performance. In comparison with the other models, the improvements of OA values achieved by SAGN are 5.72% (over AlexNet+SVM), 3.26% (over DCNN$_1$), 3.1% (over GoogleNet+SVM), 2.83% (over DCNN$_2$), 2.76% (over VGG-D+SVM), 2.6% (over MSCP$_1$), 2.57% (over MSCP+MRA$_1$), 2.15% (over Binary pattern encoded CNNs), 1.48% (over MSCP$_2$), 1.44% (over MSCP+MRA$_2$), 1.27% (over Context aggregation), 0.87% (over RTN), 0.83% (over MG-CAP with Sqrt-E), 1.36% (over DFGCN), and 0.83% (over H-GCN). The reasons why SAGN has the superior performance are two folds. First, the introduced dense feature pyramid network fuses multi-scale features, which could increase the discrimination of visual representation. Second, by using the adaptive semantic analysis and dynamic graph convolution modules, the short-/long-range semantic context information hidden in HRRS scenes can be exploited. In other words, the global and local contents of HRRS scenes can be fully captured.

2) Results on AID: The performance of all models is summarized in Table VII, and the behavior of our SAGN is the strongest in all cases. For example, when the ratio of training set equals 20%, compared with the other models, the performance enhancements achieved by SAGN are 12.99% (over AlexNet+SVM), 11.15% (over DCNN$_1$), 8.75% (over GoogleNet+SVM), 7.19% (over DCNN$_2$), 6.94% (over MSCP$_1$), 6.54% (over VGG-D+SVM), 4.99% (over MSCP+MRA$_1$), 3.99% (over MSCP$_2$), 3.21% (over MSCP+MRA$_2$), 1.96% (over MG-CAP with Sqrt-E), 1.45% (over Binary parttern encoded CNNs), 5.20% (over DFGCN), and 2.25% (over H-GCN). When the ratio of training set equals 50%, the performance of SAGN is still the best.

3) Results on NWPU: We list the OA values of compared methods and SAGN in Table VIII, which contains two columns corresponding to the training ratio with 10% and 20%. When the training set ratio equals 10%, SAGN achieves the highest performance. In comparison with the second model MG-CAP with Sqrt-E, the accuracy improvement achieves 0.99%. When the training set ratio is equal to 20%, the accuracy of SAGN obviously exceeds the listed methods (except H-GCN). H-GCN achieve the first accuracy (93.62%), which also demonstrates the superiority of GCN. By observing Table VIII, we can find that there is a distinct performance gap between two SAGNs (up to 1.76%). It means more diverse training samples the better performance, especially to the complex data set.

There are two GCN-based models in the compared methods, i.e., DFGCN [29] and H-GCN [30]. To further study the performance of our SAGN in comparison with them, we choose the other three assessment criteria, including precision, sensitivity, and specificity [60]. Also, the t-distributed stochastic neighbor embedding (t-SNE) algorithm [61] is selected to investigate the structure of features obtained by three GCN-based models. The detailed results and related discussion can be found in the supplementary material.

TABLE VIII

OVERALL ACCURACY AND STANDARD DEVIATIONS OF THE PROPOSED SAGN AND THE COMPARED METHODS ON THE NWPU DATA SET (%). THE ENTRY WITH THE HIGHEST VALUE IS SHOWN IN BOLD. (*: CNN-BASED METHOD; †: GCN-BASED METHOD)

Method	OA	
	T.R. = 10%	T.R. = 20%
AlexNet+SVM [59] *	81.22 ± 0.19	85.16 ± 0.18
MSCP$_1$ [58] *	81.70 ± 0.23	85.53 ± 0.16
GoogleNet+SVM [59] *	82.57 ± 0.12	86.02 ± 0.18
MSCP$_2$ [58] *	85.33 ± 0.17	88.93 ± 0.14
DCNN$_1$ [59] *	85.56 ± 0.20	87.24 ± 0.12
DCNN$_2$ [59] *	86.89 ± 0.10	90.49 ± 0.15
VGG-D+SVM [59] *	87.15 ± 0.45	90.36 ± 0.18
MSCP+MRA$_2$ [58] *	88.07 ± 0.18	90.81 ± 0.13
MSCP+MRA$_1$ [58] *	88.31 ± 0.23	87.05 ± 0.23
RTN [28] *	89.90	92.71
Context aggregation [40] *	90.70 ± 0.18	93.47 ± 0.26
MG-CAP with Sqrt-E [4] *	90.83 ± 0.12	92.95 ± 0.13
DFGCN [29] †	86.42 ± 0.16	89.29 ± 0.28
H-GCN [30] †	91.39 ± 0.19	**93.62 ± 0.28**
SAGN (our) †	**91.73 ± 0.18**	93.49 ± 0.10

TABLE IX

THE OA AND TC (MILLISECOND) OF THE PROPOSED SAGN AND THE COMPARED METHODS ON THE UCM AND AID DATA SET

Method	UCM (T.R. = 80%)		AID (T.R. = 50%)	
	OA ↑	TC ↓(*ms*)	OA ↑	TC ↓ (*ms*)
DFGCN [29]	98.48 ± 0.42	6.98	94.88 ± 0.22	6.54
H-GCN [30]	99.00 ± 0.43	2.89	95.78 ± 0.37	2.24
SAGN(our)	**99.82 ± 0.10**	**1.69**	**96.77 ± 0.18**	**1.42**

F. Time Consumption

In this section, we study the efficiency of our model. To this end, we record its inference time for a single HRRS image. Also, the inference time costs of two GCN-based methods, DFGCN [29] and H-GCN [30], are counted for reference. The results are listed in Table IX, which consists of OA and time costs (TC). Here, the higher OA is (↑) and the fewer TC is (↓), the better performance. As shown in Table IX, SAGN achieves the highest accuracy with the lowest time consumption, which shows the superiorities of our model again.

V. CONCLUSION

In this paper, an end-to-end HRRS image scene classification network (SAGN) has been proposed with the consideration of complex contents within HRRS images. Instead of only taking the single semantic label into account, SAGN exploits various semantics from HRRS scenes and mines their unstructured relations to decide the final semantic categories. For an HRRS scene, it is mapped into feature maps by a dense feature pyramid network first. This step helps capture the multi-scale information from the HRRS scene. Second, an adaptive semantic analysis module has been conducted to analyze the HRRS scene into different regions using multi-scale feature maps. These regions correspond to diverse semantics so that the contents that are not described by the manual annotation can be explored. Third, to study the relationships between different semantic regions, a learnable adjacency matrix construction method has been developed. Then, under the GCN paradigm, a dynamic graph feature update module has been established to learn the pixel-level features using the regional adjacency matrix. Therefore, the representation of semantic regions and the pixels can be updated simultaneously, which helps improve the discrimination of the learned features. Finally, the category of the HRRS scene is decided by all of the semantic regions. The positive experimental results on three public data sets (i.e., NWPU, AID, and UCM) have demonstrated the efficacy of SAGN on HRRS image scene classification tasks.

ACKNOWLEDGMENT

The authors would like to thank Prof. Handing Wang, School of Artificial Intelligence, Xidian University, Xi'an, China, for her assistance in the title.

REFERENCES

[1] Y. Guo, X. Jia, and D. Paull, "Effective sequential classifier training for SVM-based multitemporal remote sensing image classification," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 3036–3048, Jun. 2018.

[2] L. Chen, W. Zhan, W. Tian, Y. He, and Q. Zou, "Deep integration: A multi-label architecture for road scene recognition," *IEEE Trans. Image Process.*, vol. 28, no. 10, pp. 4883–4898, May 2019.

[3] G. Chen, X. Song, H. Zeng, and S. Jiang, "Scene recognition with prototype-agnostic scene layout," *IEEE Trans. Image Process.*, vol. 29, pp. 5877–5888, 2020.

[4] S. Wang, Y. Guan, and L. Shao, "Multi-granularity canonical appearance pooling for remote sensing scene classification," *IEEE Trans. Image Process.*, vol. 29, pp. 5396–5407, 2020.

[5] L. Wang, S. Guo, W. Huang, Y. Xiong, and Y. Qiao, "Knowledge guided disambiguation for large-scale scene classification with multi-resolution CNNs," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 2055–2068, Apr. 2017.

[6] L. Du and H. Ling, "Dynamic scene classification using redundant spatial scenelets," *IEEE Trans. Cybern.*, vol. 46, no. 9, pp. 2156–2165, Sep. 2016.

[7] S. E. Grigorescu, N. Petkov, and P. Kruizinga, "Comparison of texture features based on Gabor filters," *IEEE Trans. Image Process.*, vol. 11, no. 10, pp. 1160–1167, Oct. 2002.

[8] X. Tang, L. Jiao, and W. J. Emery, "SAR image content retrieval based on fuzzy similarity and relevance feedback," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 5, pp. 1824–1842, May 2017.

[9] S. Mei, J. Ji, J. Hou, X. Li, and Q. Du, "Learning sensor-specific spatial–spectral features of hyperspectral images via convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4520–4533, Aug. 2017.

[10] L. Jiao, X. Tang, B. Hou, and S. Wang, "SAR images retrieval based on semantic classification and region-based similarity measure for Earth observation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 8, pp. 3876–3891, Aug. 2015.

[11] S. Sergyan, "Color histogram features based image classification in content-based image retrieval systems," in *Proc. 6th Int. Symp. Appl. Mach. Intell. Informat.*, Jan. 2008, pp. 221–224.

[12] X. Tang and L. Jiao, "Fusion similarity-based reranking for SAR image retrieval," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 2, pp. 242–246, Feb. 2017.

[13] H. Soltanian-Zadeh, F. Rafiee-Rad, and S. Pourabdollah-Nejad, "Comparison of multiwavelet, wavelet, haralick, and shape features for microcalcification classification in mammograms," *Pattern Recognit.*, vol. 37, no. 10, pp. 1973–1986, Oct. 2004.

[14] X. Tang, L. Jiao, W. J. Emery, F. Liu, and D. Zhang, "Two-stage reranking for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 10, pp. 5798–5817, Oct. 2017.

[15] L. Zhang, W. Zhou, and L. Jiao, "Wavelet support vector machine," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 34, no. 1, pp. 34–39, Feb. 2004.

[16] M. A. Friedl and C. E. Brodley, "Decision tree classification of land cover from remotely sensed data," *Remote Sens. Environ.*, vol. 61, pp. 399–409, Sep. 1997.

[17] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.*, vol. 53, pp. 5455–5516, Apr. 2020.

[18] J. Song, S. Gao, Y. Zhu, and C. Ma, "A survey of remote sensing image classification based on CNNs," *Big Earth Data*, vol. 3, no. 3, pp. 232–254, Jul. 2019.

[19] C. Liu, J. Ma, X. Tang, F. Liu, X. Zhang, and L. Jiao, "Deep hash learning for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3420–3443, Apr. 2021.

[20] X. Tang et al., "Hyperspectral image classification based on 3-D octave convolution with spatial–spectral attention network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2430–2447, Mar. 2020.

[21] J. Ma, L. Wu, X. Tang, F. Liu, X. Zhang, and L. Jiao, "Building extraction of aerial images by a global and multi-scale encoder–decoder network," *Remote Sens.*, vol. 12, no. 15, p. 2350, Jul. 2020.

[22] X. Tang, C. Liu, J. Ma, X. Zhang, F. Liu, and L. Jiao, "Large-scale remote sensing image retrieval based on semi-supervised adversarial hashing," *Remote Sens.*, vol. 11, no. 17, p. 2055, 2019.

[23] X. Tang, X. Zhang, F. Liu, and L. Jiao, "Unsupervised deep feature learning for remote sensing image retrieval," *Remote Sens.*, vol. 10, no. 8, p. 1243, Aug. 2018.

[24] Y. Zhao et al., "Hyperspectral image classification via spatial window-based multiview intact feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2294–2306, Mar. 2021.

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

[26] Y. Yu, X. Li, and F. Liu, "Attention GANs: Unsupervised deep feature learning for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 519–531, Jan. 2020.

[27] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2018.

[28] L. Fu, D. Zhang, and Q. Ye, "Recurrent thrifty attention network for remote sensing scene recognition," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 10, pp. 8257–8268, Oct. 2021.

[29] K. Xu, H. Huang, P. Deng, and Y. Li, "Deep feature aggregation framework driven by graph convolutional network for scene classification in remote sensing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5751–5765, Oct. 2021.

[30] Y. Gao, J. Shi, J. Li, and R. Wang, "Remote sensing scene classification based on high-order graph convolutional network," *Eur. J. Remote Sens.*, vol. 54, no. 1, pp. 141–155, Feb. 2021.

[31] F. Manessi, A. Rozza, and M. Manzo, "Dynamic graph convolutional networks," *Pattern Recognit.*, vol. 97, Jan. 2020, Art. no. 107000.

[32] X. Yang et al., "Automatic ship detection in remote sensing images from Google earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sens.*, vol. 10, no. 1, p. 132, 2018.

[33] Y. Gu, Y. Wang, and Y. Li, "A survey on deep learning-driven remote sensing image scene understanding: Scene classification, scene retrieval and scene-guided object detection," *Appl. Sci.*, vol. 9, no. 10, p. 2110, 2019.

[34] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning earth observation classification using ImageNet pretrained networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, Jan. 2015.

[35] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "OverFeat: Integrated recognition, localization and detection using convolutional networks," 2013, *arXiv:1312.6229*.

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

[37] Y. Liu, Y. Zhong, and Q. Qin, "Scene classification based on multiscale convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7109–7121, Dec. 2018.

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

[39] X. Zheng, Y. Yuan, and X. Lu, "A deep scene representation for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4799–4809, Jul. 2019.

[40] Z. Zhou, Y. Zheng, H. Ye, J. Pu, and G. Sun, "Satellite image scene classification via convnet with context aggregation," in *Proc. Pacific Rim Conf. Multimedia*. Hefei, China: Springer, 2018, pp. 329–339.

[41] X. Peng and A. Bouzerdoum, "Part-based feature aggregation method for dynamic scene recognition," in *Proc. Digital Image Comput. Techn. Appl. (DICTA)*, Dec. 2019, pp. 1–8.

[42] L. He, J. Bai, and M. Yang, "Feature aggregation convolution network for haze removal," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 2806–2810.

[43] X. Lu, H. Sun, and X. Zheng, "A feature aggregation convolutional neural network for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7894–7906, Oct. 2019.

[44] R. M. Anwer, F. S. Khan, J. van de Weijer, M. Molinier, and J. Laaksonen, "Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification," *ISPRS J. Photogramm. Remote Sens.*, vol. 138, pp. 74–85, Apr. 2018.

[45] Y. Li, Y. Zhang, and Z. Zhu, "Error-tolerant deep learning for remote sensing image scene classification," *IEEE Trans. Cybern.*, vol. 51, no. 4, pp. 1756–1768, Apr. 2021.

[46] S. Chaudhari, V. Mithal, G. Polatkan, and R. Ramanath, "An attentive survey of attention models," 2019, *arXiv:1904.02874*.

[47] X. Tang, Q. Ma, X. Zhang, F. Liu, J. Ma, and L. Jiao, "Attention consistent network for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2030–2045, 2021.

[48] M. Steyvers and T. Griffiths, "Probabilistic topic models," *Handbook Latent Semantic Anal.*, vol. 427, no. 7, pp. 424–440, 2007.

[49] Q. Zhu, Y. Zhong, L. Zhang, and D. Li, "Adaptive deep sparse semantic modeling framework for high spatial resolution image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 6180–6195, Oct. 2018.

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

[51] J. Zhang, J. Liu, B. Pan, and Z. Shi, "Domain adaptation based on correlation subspace dynamic distribution alignment for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7920–7930, Nov. 2020.

[52] S. Wan, C. Gong, P. Zhong, B. Du, L. Zhang, and J. Yang, "Multiscale dynamic graph convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3162–3177, May 2020.

[53] L. Mou, X. Lu, X. Li, and X. X. Zhu, "Nonlocal graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8246–8257, Dec. 2020.

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

[55] G.-S. Xia et al., "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

[56] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[57] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," 2019, *arXiv:1912.01703*.

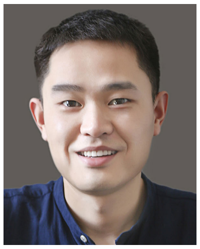
[58] N. He, L. Fang, S. Li, A. Plaza, and J. Plaza, "Remote sensing scene classification using multilayer stacked covariance pooling," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 6899–6910, Dec. 2018.

[59] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[60] S. Dua, U. R. Acharya, P. Chowriappa, and S. V. Sree, "Wavelet-based energy features for glaucomatous image classification," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 1, pp. 80–87, Jan. 2012.

[61] A. C. Belkina, C. O. Ciccolella, R. Anno, R. Halpert, J. Spidlen, and J. E. Snyder-Cappione, "Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets," *Nature Commun.*, vol. 10, no. 1, pp. 1–12, Nov. 2019.

Yuqun Yang (Student Member, IEEE) received the B.Sc. degree in information and computing science from the Xi'an University of Technology, Xi'an, China, in 2019. He is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Xidian University, Xi'an. His research interests include machine learning, object detection, and image classification.



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

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

Yiu-Ming Cheung (Fellow, IEEE) received the Ph.D. degree from the Department of Computer Science and Engineering, Chinese University of Hong Kong, Hong Kong. He is currently a Chair Professor (Artificial Intelligence) with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His research interests include machine learning and visual computing, and their applications. He is a fellow of AAAS, IET, and BCS. He is the Editor-in-Chief of the IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE. He also serves as an Associate Editor for the IEEE TRANSACTIONS ON CYBERNETICS, IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS, *Pattern Recognition*, and *Neurocomputing*. More details can be found at: https://www.comp.hkbu.edu.hk/~ymc

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

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

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

From 1984 to 1986, he was an Assistant Professor with the Civil Aviation Institute of China, Tianjin, China. From 1990 to 1991, he was a Postdoctoral Fellow with the Key Laboratory for Radar Signal Processing, Xidian University, Xi'an, where he is currently the Director of the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education. He has authored or coauthored more than 200 scientific articles. His research interests include signal and image processing, nonlinear circuits and systems theory, wavelet theory, natural computation, and intelligent information processing. He is a member of the IEEE Xi'an Section Executive Committee and an Executive Committee Member of the Chinese Association of Artificial Intelligence. He is also the Chairperson of the Awards and Recognition Committee.