Zero-Shot Remote Sensing Scene Classification Method Based on Local-Global Feature Fusion and Weight Mapping Loss

Chao Wang, Junyong Li, Ahmed Tanvir, Jiajun Yang, Tao Xie, Liqiang Ji, and Tong Zhang

***Abstract*—Zero-shot remote sensing scene classification refers to making the model to have the ability to identify the unseen class scenes based on seen class scenes, and has become a research hotspot in the field of remote sensing. Contemporary approaches in zero-shot remote sensing scene classification primarily focus on extracting global information from scenes, neglecting nuanced local landscape features. This oversight diminishes the discriminative capabilities of recognition models. Furthermore, these methods overlook the semantic relevance between seen and unseen class scenes in training, leading to reduced emphasis on learning from varied scenes and subsequent declines in classification performance. To address these challenges, this article proposes the "Zero-Shot Remote Sensing Scene Classification Method Based on Local-Global Feature Fusion and Weight Mapping Loss (LGFFWM)." The design incorporates a local-global feature fusion (LGFF) module enabling adaptive labeling and feature modeling of internal local landscapes, effectively merging them with global features for a more discriminative representation of remote sensing scenes. Furthermore, a weight mapping loss (WM Loss) function is introduced, leveraging a semantic correlation matrix to compel the model to prioritize learning seen class scenes that exhibit strong correlations with unseen class scenes by assigning higher training weights. Extensive experiments have been conducted on classical remote sensing scene datasets, including UCM, AID, and NWPU, demonstrate the superiority of the proposed LGFFWM method over ten advanced comparative methods, yielding overall accuracy improvements of over 2.25%, 3.47%, and 0.44%, respectively. Additional experiments on the SIRI-WHU and RSSCN7 datasets underscore the transferability of LGFFWM, achieving overall accuracies of 53.50% and 47.37%, respectively.**

***Index Terms*—Local-global feature fusion (LGFF), remote sensing scene classification (RSSC), weight mapping, zero-shot learning (ZSL).**

Manuscript received 7 September 2023; revised 16 November 2023; accepted 7 December 2023. Date of publication 21 December 2023; date of current version 12 January 2024. This work was supported in part by the Post-Doctoral Fund of Jiangsu Province under Grant 2021K013A, in part by the National Key Research and Development Program of China under Grant 2022YFC3004202, and in part by the National Natural Science Foundation of China under Grant 42176180. *(Corresponding author: Tao Xie.)*

Chao Wang, Junyong Li, Ahmed Tanvir, Jiajun Yang, Liqiang Ji, and Tong Zhang are with the School of Electronics and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: chaowang@nuist.edu.cn; 20211249174@nuist.edu.cn; 201953050001@nuist.edu.cn; 202212490674@nuist.edu.cn; 202212490153@nuist.edu.cn; 202312490210@nuist.edu.cn).

Tao Xie is with the School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: xtplqk@126.com).

Digital Object Identifier 10.1109/JSTARS.2023.3344628

I. Introduction

REMOTE sensing images have the advantages of fast acquisition speed and wide coverage [1], [2], [3]. Based on high-resolution remote sensing (HRS) images, remote sensing scene classification (RSSC) is of great significance for prompt and accurately obtaining urban development information and scientifically conducting urban planning [4], [5], [6]. Unlike traditional object classification tasks, remote sensing scenes are highly abstract semantic concepts that do not always correspond to a specific land feature type. Therefore, traditional pixel-level or object-level classification methods [7], [8], [9] are difficult to apply directly. Currently, HRS image classification for remote sensing scenes has become a research hotspot in the fields of remote sensing and computer vision.

In order to construct a more discriminating features to describe remote sensing scenes, researchers have carried out extensive research work and achieved many achievements [10], [11], [12], [13], [14], [15]. Among them, the descriptors based on artificial design rely on prior knowledge and have the advantages of low computation and strong interpretability. For example, Zhu et al. [10] proposed a knowledge-guided land pattern description framework, which uses adaptive gradient perception mechanism and land pattern cognitive model to capture the internal and external relations between different land cover types. Lv et al. [13] proposed an adaptive region-based Gauss-weighted spectral (GWS) to improve the spectral homogeneity of local regions around pixels. On this basis, an area shape index (ASI), which describes the relationship between the area and shape of the adaptive region around each pixel is proposed and combined with GWS to address the problem of insufficient basic image features for land cover classification. Compared with artificially designed descriptors, the method based on deep learning gets rid of the dependence on prior knowledge and can automatically extract highly abstract and highly discriminating features, which has become the main technical means for scene classification [14], [16], [17]. For example, Penatti et al. [18] evaluated the generalization ability of deep features (ConvNets) in two new scenes, aerial and remote sensing images, for the classification of aerial and remote sensing images. Zhu et al. [11] proposed a weakly pseudosupervised decorrelated subdomain adaptation (WPS-DSA) framework for high spatial resolution cross-domain land use classification of high-speed rail. Chaib et al. [19] used visual geometry group network (VGGNet) as a feature

extractor to select more representative deep features to optimize the representation of scenes. At the same time, although these methods have shown good performance in specific datasets, they still have prominent limitations in practical applications, mainly manifested in the following: 1) due to factors, such as economic development level and cultural differences, the visual features of the same urban remote sensing scene may vary greatly in different countries or even different regions of the same country [20]. Therefore, the classification models trained on specific datasets are usually not transferable, and it is difficult to directly apply them to the classification of remote sensing scenes in different cities and 2) modern urban remote sensing scenes are diverse and in continuous evolution [21], [22], while existing models can only classify the limited number of labeled scene categories provided in the training set. For other unlabeled or newly emerging scene categories, the models do not have classification capabilities, i.e., the generalization ability and scalability of the models are poor.

In order to cope with the above challenges, some researchers proposed few-shot learning (FSL) and zero-shot learning (ZSL) for RSSC. The former reduces the requirement on the number of scenes for model training. For example, Huang et al. [23] proposed a task-adaptive embedding network, TAE-Net, to enhance the generalization ability of the model for unseen remote sensing scenes under the condition of FSL. Wang et al. [24] proposed a novel transductive learning framework with conditional metric embedment to deal with the problems of interclass metric bias and intraclass variation under the condition of FSL. Meanwhile, a transductive prototype learning strategy was proposed to enhance the robustness of prototypes to intra-class variation. Zeng et al. [25] proposed a calibration method for scene classification prototype of remote sensing scenes with FSL based on feature generation model combined with self-attention feature encoder. Compared with FSL, ZSL completely gets rid of the dependence on seen class scenes for RSSC tasks. ZSL can use the semantic relevance between seen and unseen classes as a bridge to make the classification model have good transferability and the ability to classify unseen class scenes. Therefore, compared with FSL, it has better usability in practical applications. For example, Li et al. [26] proposed a novel remote sensing knowledge graph (RSKG), which fully considers the rich connections between remote sensing scene categories. Generate a semantic representation of scene categories by representation learning of RSKG (SR-RSKG). Wang et al. [27] proposed a distance-constrained semantic autoencoder to reduce the semantic gap between visual features and semantic representations, which to some extent alleviates the domain shift problem. Quan et al. [28] designed a semisupervised Sammon embedding algorithm to transfer the unseen knowledge in the semantic space to the visual space, making it more consistent with the class structure in the visual space. Wu et al. [29] proposed a transductive zero-shot RSSC algorithm based on Sammon embedding to address the problem of inconsistency between scene classes in semantic space and visual space and the domain shift problem. The authors in [30] constructed a semantic directed graph to describe the relationship between seen and unseen classes, and used a label propagation algorithm for zero-shot classification. Li et al. [31] used generative adversarial networks for zero-shot RSSC. Ma et al. [32] proposed to use generative adversarial networks (GANs) to enhance the variational autoencoder generative model to better measure the reconstruction quality in zero-shot RSSC.

However, the above ZSL methods still have significant limitations and shortcomings in the following two aspects: 1) they only consider global information while constructing the visual space and ignore the representative local landscape in different urban remote sensing scenes; on the other hand, the typical ground objects and their spatial distribution information contained in those local landscapes are often very discriminative for scene classification tasks [33], [34], [35]. Therefore, it is necessary to further introduce local landscape features to obtain a more discriminating scene representation. However, at present, local landscapes corresponding to different scenes are usually obtained based on manual annotation, such as knowledge graph [26], [36]. These methods not only have a low level of automation, but also may have differences in the prior knowledge of annotators, resulting in poor reliability of annotation results. So, it is urgent to find a method that can adapt to extract local landscape representation according to the characteristics of the scene itself and 2) these methods do not consider the difference in semantic relevance between each seen class and unseen class scenes in the training set [37], [38], [39], [40]. In fact, the model should pay more attention to the learning of unseen class scenes with a higher semantic relevance to seen classes, so as to enhance the classification ability of unseen class scenes. Therefore, this article assigns different weights to each seen class scene in the training, so as to more fully mine the relevant information between the seen class and the unseen class scenes. Based on the above analysis, this article proposes the "Zero-Shot Remote Sensing Scene Classification Method Based on Local-Global Feature Fusion and Weight Mapping Loss (LGFFWM)." The main contributions and contributions are as follows.

1) In order to fully exploit the complementary advantages of global and local landscapes in RSSC tasks, this article designed a local-global feature fusion (LGFF) module. Different from the traditional manual labeling method, LGFF can start from the scene itself, adaptively extracts the set of proposal frames for marking local landscape according to the type and spatial distribution of ground objects and use the proposed normalized local feature matrix is used to model the feature of the local landscape. Furthermore, the local and global features are effectively integrated by cascading method, so as to achieve more discriminating remote sensing scene representation.
2) This article proposes a weight mapping loss (WM Loss) function based on semantic correlation matrix to more fully mine relevant semantic information to enhance the classification ability of the model under zero-shot conditions. WM Loss utilizes a normalized semantic correlation matrix to evaluate the semantic relevance between each seen class and each unseen class scene. On this basis, the loss function can force the model to assign higher weight to the seen class scenes with better correlation with the unseen class scenes in the training phase, so as to force the model to strengthen the learning of such scenes.

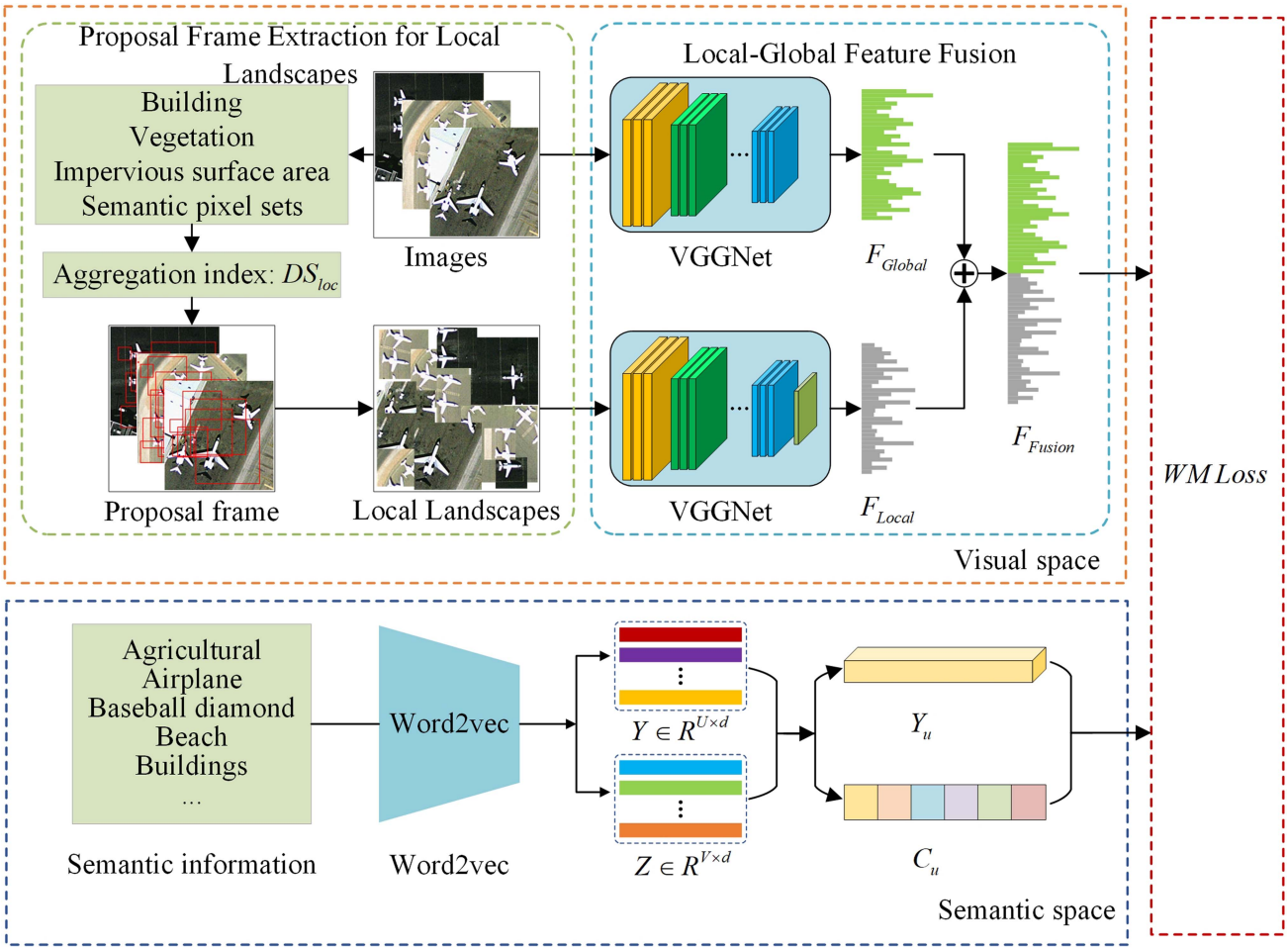


Fig. 1. LGFFWM network architecture.

This article mainly includes five sections. The next section introduces the proposed method in detail; then the experimental settings and results are explained and discussed; the last section concludes the contribution of the article.

II. Methodology

This section provides a detailed explanation of the proposed method. First, it provides an overview of the zero-shot classification framework established in the article. Based on this, it further explains the LGFF module. Finally, it provides the working principle of the proposed WM Loss function based on the semantic correlation matrices.

A. Model Overview

The architecture of the zero-shot RSSC model proposed in this article is depicted in Fig. 1, consisting primarily of visual space and semantic space, with visual space working as the embedding space [41]. In the deep convolutional neural network architecture, VGGNet can obtain a score vector by connecting a softmax layer after the backbone network [42], facilitating local landscape selection. Through the deep convolutional neural network architecture, VGGNet effectively extracts high-level and low-level features [43], [44], [45]. These high-level features are then combined through fully connected layers. Therefore, VGGNet is adopted as the base network. In the visual space, starting from the original image, the set of proposal frames are used for marking local landscapes is performed, followed by feature modeling. Simultaneously, effective fusion of local and global features is achieved through a cascaded approach [46], [47], resulting in a more distinguished representation of scenes. In the semantic space, the Word2Vec [48] model trained on the Wikipedia corpus was used to transform each scene class into a semantic vector, which is then mapped to the visual space. The loss function adopts the WM Loss proposed by the text and is employed to measure the differences in semantic correlation between different scene categories, enabling a more comprehensive exploration of relevant semantic knowledge.

B. LGFF Module

The LGFF module, as conceived in this article, primarily realizes the adaptive extraction of proposal frames for marking local landscape and feature modeling, in addition to the integration of local-global features. Subsequent sections will expound on the comprehensive workflow of LGFF.

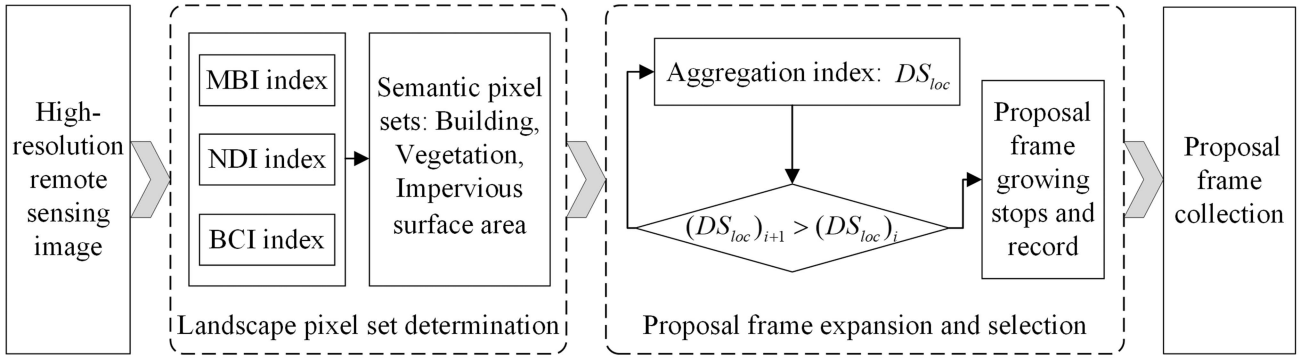


Fig. 2. Framework of proposal frame extraction for local landscape annotation.

1) Proposal Frame Extraction for Local Landscapes: To annotate representative local landscapes in urban remote sensing scenes, a novel proposal frame extraction strategy is proposed, different from traditional methods as, non-maximum suppression [49]. The strategy is designed to align with the spatial distribution and types of land cover within the scene. In urban remote sensing scenes, it is often the case that the denser and more diverse the distribution of land features, the more likely a proposal frame represents a local landscape. Based on this assumption, the workflow for proposal frame extraction is illustrated in Fig. 2.

Given the diverse types and manifestations of typical local urban landscapes, such as residential areas, commercial zones, parks, etc. For this purpose, this article selects (including, but not limited to) three representative land feature types: 1) buildings; 2) vegetation; and 3) impervious surfaces, and use MBI index [50], NDI index [51], and BCI index [52] to extract the corresponding semantic pixel set. On this basis, this article proposes a clustering index DS_{loc}, which combines the richness of feature types and their distribution density, and adaptively extracts the set of proposal frames for marking the local landscape from the original image.

$$D\,S_{loc} = \frac{\sum B_m \left(1 - \frac{di}{H}\right)}{\sum \left(1 - \frac{di}{H}\right)}. \tag{1}$$

Here, if the current proposal frame is denoted as $Frame$, H represents the diagonal length of $Frame$, di is the distance from a particular $pixel$ to the center pixel within $Frame$, and $B_m(m = 1, 2, 3)$ represents the values of MBI, NDI, and BCI for a given pixel, respectively. Based on this, the specific steps for proposal frame generation are as follows.

Step 1: For a particular pixel, $pixel_i$ in the image, use the pixel as the center and perform proposal frame growth with 8-CONNECTIVITY [53]. Record the DS_{loc} after growing each proposal frame. When DS_{loc} increases continuously for three iterations and then. decreases continuously for three iterations, proposal frame growing stops and record the current proposal frame as $Frame_i$. Otherwise, if the boundary is reached during proposal frame growth, there is no corresponding proposal frame for $pixel_i$.

Step 2: Repeat Step 1 by traversing all pixels in the image to obtain the proposal frame collection $Frame_{all}$.

Step 3: To avoid falling into local optima, let the total number of image pixels be $pixel_{tot}$ and divide it into ten equal intervals. On this basis, count the number of pixels in each proposal frame in $Frame_{all}$ and cluster them based on the respective intervals. Finally, select the proposal frames with the highest DS_{loc} value (or tied highest values) from each interval to form the final proposal frame collection $Frame_{opt}$ for each image x_i.

2) LGFF: Distinct local landscapes in urban remote sensing scenes are primarily characterized by significant differences in the distribution of land feature, while each local landscape belonging to the similar kind of scene typically corresponding to a specific land feature or a combination of land features that appear with higher frequency.

Therefore, based on $Frame_{opt}$, this article defines a normalized local feature matrix to describe local landscapes. Initially, VGGNet is pretrained on the ImageNet dataset to enable the model to classify 1000 classes of objects, including typical urban targets, such as buildings, trains, and ponds. Subsequently, for a proposal frame corresponding to x_i, it is input into VGGNet to obtain a 1000-dimensional score vector $\boldsymbol{S}_q$ from the softmax layer output. Each dimension in $\boldsymbol{S}_q$ corresponds to a class of object o in the ImageNet dataset and reflects the likelihood of the presence of o within the current proposal frame. By iterating through all Q proposal frames in $Frame_{opt}$, the local feature matrix $\boldsymbol{S} = [\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_Q]$ corresponding to x_i is obtained. Taking into consideration that the number of proposal frames contained in each image may vary, the normalized local feature matrix F_{Local} for x_i is further defined as follows:

$$F_{Local} = \frac{1}{Q} \sum_{q=1}^{Q} \boldsymbol{S}_q. \tag{2}$$

Furthermore, the output of the fully connected layers of VGGNet is employed as the global feature F_{Global} for x_i. Subsequently, local and global features are fused through a cascading approach to obtain the visual feature F_{Fusion}.

$$F_{Fusion} = [F_{Global}, F_{Local}]. \tag{3}$$

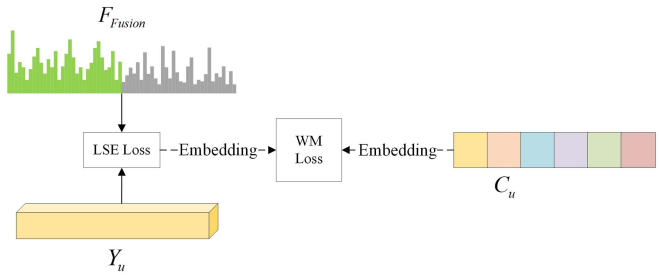


Fig. 3. Illustration of the WM loss.

C. Weight Mapping Loss

To encourage the model to focus more on learning from seen class samples that are more relevant to unseen classes, at first, semantic vector sets $\boldsymbol{Y} = \{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_U\}$ and $\boldsymbol{Z} = \{\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_V\}$ for seen and unseen scene classes are obtained based on Word2Vec, respectively. Here, U and V represent the total number of categories for seen and unseen scene classes, respectively.

For a seen class scene class u, a normalized semantic relevance measure C_u is constructed using a Gaussian kernel function, as follows:

$$C_u = \frac{1}{V} \sum_{v=1}^{V} e^{-\|\boldsymbol{Y}_u - \boldsymbol{Z}_v\|^2}. \tag{4}$$

As it can be seen that, C_u reflects the average semantic relevance degree between u and all unknown scene classes. Furthermore, this article builds the proposed WM Loss by embedding C_u based on the popular LSE Loss (least square embedding loss) [16]. The LSE Loss connects two branches together, significantly reducing the disparity between visual features and semantic features in the visual space. The formula for calculating WM Loss is as follows:

$$\mathcal{L}\ (u) = C_u \cdot \sum ||F_{Fusion} - \mathbf{Y}_{\boldsymbol{u}}||^2 + \lambda ||\boldsymbol{W}||^2 \tag{5}$$

where $\boldsymbol{W}$ is a randomly initialized encoding matrix for the semantic embedding branch, aiming to align the visual space and semantic space, and λ is a regularization parameter. Fig. 3 illustrates the workflow of the proposed WM Loss.

III. Experiments and Analysis of Results

In order to fully evaluate the performance of the proposed LGFFWM method, three representative datasets were selected for the experiments and then, results were compared with ten state-of-the-art methods.

A. Datasets

The experiments were conducted on three benchmark remote sensing scene datasets. They are UC Merced (UCM) [54], the Aerial Image Dataset (AID) [55], and NWPU-RESISC45 (NWPU) [56]. UCM is derived from the United States Geological Survey National Map Urban Area Imagery series, providing labeled samples of various categories of typical urban remote sensing scenes. AID, released by the Huazhong University of Science and Technology and Wuhan University, is a large-scale aerial image dataset constructed from samples collected from Google Maps images. NWPU, published by Northwestern Polytechnical University, is an open dataset with significant variations among different scene samples in terms of translation, spatial resolution, and more. Using UCM, AID, and NWPU datasets allows for a comprehensive analysis of performance the proposed method from different perspectives. Moreover, to substantiate the transferability of LGFFWM, experiments executed on the SIRI-WHU [57] and RSSCN7 [58] datasets. Among them SIRI-WHU, released by Wuhan University, offers high-resolution satellite images with resolutions reaching up to 2 m. RSSCN7, an openly accessible dataset also originating from Wuhan University, encompasses images captured across diverse seasons and weather conditions, thereby introducing notable challenges. Comprehensive parameter comparisons for each dataset are presented in Table I.

B. Methods for Comparison and Experimental Settings

1) Comparison With State-of-the-Art Methods: In order to thoroughly evaluate the performance of the LGFFWM method, we have selected ten different state-of-the-art methods for comparison. These baseline models include SSE [37], DMaP [38], SAE [39], ZSL-LP [30], ZSC-SA [28], VSOP [59], f-CLSWGAN [60], CYCLEWGAN [61], RBGN [40], and DSAE [27].

Zhang and Saligrama [37] mapped source domain and target domain data into the same semantic space and calculated their relevance. Li et al. [38] explored the intrinsic relationship between semantic space manifolds and the transfer ability of visual semantic mappings. Li et al. [30] constructed a semantic directed graph to describe the relationship between seen and unseen classes. Gaussian kernel weighted distance is used to establish directed edges and Google Inception Net (GoogleNet) network for feature extraction. Finally, employs a label propagation algorithm for zero-shot classification. Kodirov et al. [39] is based on a learning semantic autoencoder that projects visual features into semantic space and reconstructs the original visual features. SAE adds constraints from visual mapping to semantic features to mitigate domain shift problems. Wang et al. [27] proposed a distance-constrained semantic autoencoder to process zero-shot RSSC. Quan et al. [28] adopted a semisupervised Sammon embedding algorithm to transfer unseen knowledge from the semantic space to the visual space, making it consistent with the class structure of the visual space prototypes. Wu et al. [59] proposed a new approach to guide visual semantic embeddedness learning by using the mutual information between visual and semantic features. Xian et al. [60] proposed a new generative adversarial network (GAN) for ZSL. Felix et al. [61] proposed a new regularization based on GAN training, which uses this constraint to force generated visual features to reconstruct their original semantic features. Based on conditional generative adversarial network of generalized zero-shot learning (GZSL), robust bidirectional generative network (RBGN) [40] proposes a novel generative method called RBGN. These methods were

TABLE I
COMPARISON OF EXPERIMENTAL DATASETS

Dataset	Number of scenes	Number of samples	Samples per class	Pixel resolution (m)	Data format	Year
UCM	21	2100	100	0.3	256×256	2010
AID	30	10 000	220-420	0.5-0.8	600×600	2016
NWPU	45	31 500	700	0.2-30	256×256	2017
SIRI-WHU	12	2400	200	2	200×200	2016
RSSCN7	7	2800	400	-	400×400	2015

TABLE II
DIVISION OF SEEN/UNSEEN CLASSES AND TRAINING/TESTING SAMPLES FOR THE SCENE DATASETS

Dataset	Seen/Unseen classes, training/testing samples				
UCM	seen/unseen	16/5	13/8	10/11	7/14
	training/testing	1600/500	1300/800	1000/1100	700/1400
AID	seen/unseen	25/5	20/10	15/15	10/20
	training/testing	8230/1770	6600/3400	4700/5300	3160/6840
NWPU	seen/unseen	35/10	30/15	25/20	20/25
	training/testing	24 500/7000	21 000/10 500	17 500/14 000	14 000/17 500
SIRI-WHU	seen/unseen	9/3	7/5	5/7	3/9
	training/testing	1800/600	1400/1000	1000/1400	600/1800
RSSCN7	seen/unseen	5/2	4/3	3/4	2/5
	training/testing	2000/800	1600/1200	1200/1600	800/2000

designed with specific considerations in mapping space, domain shift, visual and semantic feature extraction, generative adversarial network, utilizing different loss functions. Comparing our method to these baselines helps provide a comprehensive and objective evaluation of LGFFWM.

2) Experimental Settings: In terms of experimental settings, our proposed method is implemented on the Ubuntu 16.04 system, using the PyTorch-1.3.1 framework, and runs on hardware with an Nvidia GeForce RTX 2080ti GPU with 11GB of RAM. In the visual space, we have used a pretrained VGGNet on ImageNet to obtain local landscape representations, and the fully connected features for the entire image are also obtained using the pretrained VGGNet model on ImageNet. We extracted 300-dimensional word vectors using the Word2Vec model, with parameters for KNN nearest neighbor relationships set based on [41]. Additionally, the fully connected layers of the embedding space model were initialized with random weights. The learning rate for the Adam optimizer was set to 0.00001, and the minibatch size of 64.

Wang et al. [27] leveraged deep features with superior foreground classification performance compared to handcrafted features. Hence, we utilized the ResNet152 (Residual Network 152) model pretrained on the Places dataset to extract 2048-dimensional deep features as the visual representation of the scene images. For semantic features, a 300-dimensional vector obtained from a Word2Vec [48] model trained on the Wikipedia corpus was employed to represent the semantic features of scene classes. Finally, following the recommendations in [27], the experiments used four different ratios for the split between seen and unseen classes for UCM, AID, and NWPU datasets, with no overlap between seen and unseen classes. The specific seen/unseen splits and the ratio of training/testing samples are shown in Table II.

C. Results Comparison and Analysis

1) Evaluation Metrics: To quantitatively evaluate the performance of our method, we have used several quantitative evaluation metrics including overall accuracy (OA), standard deviation (SD), confusion matrix (CM), and class average accuracy (AA) [27], [62].

OA is a direct measure of the classification accuracy of the model on the entire dataset

$$OA = \frac{N_t}{N_t + N_f} \tag{6}$$

where N_t and N_f represents the number of correctly classified and misclassified samples, respectively; SD reflects the dispersion of OA; CM is a matrix whose rows and columns describe the predicted and actual classes of the samples, allowing for a detailed analysis of classification errors in different categories; AA reflects the average classification accuracy of the model for various scene classes

$$AA = \frac{1}{n} \sum_{i=1}^{n} \frac{C_i}{N_i} \tag{7}$$

where C_i is the number of samples correctly classified for class i, N_i is the total number of samples for class i, and n is the number of classes.

2) Experimental Results Analysis: Based on three datasets, the overall accuracies of our method and the comparative methods are presented in Tables III–V.

TABLE III
OA (%) AND SD (%) OF LGFFWM METHOD AND STATE-OF-THE-ART METHODS ON THE UCM

Method	16/5	13/8	10/11	7/14
SSE [37]	35.59 ± 5.90	23.42 ± 3.81	17.07 ± 3.56	10.82 ± 2.10
DMaP [38]	48.92 ± 8.71	30.91 ± 4.77	22.99 ± 4.81	17.30 ± 3.04
SAE [39]	49.50 ± 8.42	32.71 ± 6.49	24.04 ± 4.36	18.63 ± 2.76
ZSL-LP [30]	49.01 ± 8.85	31.26 ± 5.09	23.28 ± 4.13	17.55 ± 2.90
ZSC-SA [28]	50.42 ± 8.84	34.12 ± 6.10	24.68 ± 4.22	18.38 ± 2.74
VSOP [59]	46.48 ± 7.83	29.81 ± 4.56	21.97 ± 4.11	16.14 ± 2.59
f-CLSWGAN [60]	56.97 ± 11.06	36.47 ± 6.28	27.89 ± 4.99	19.34 ± 3.96
CYCLEWGAN [61]	58.36 ± 10.04	36.81 ± 5.53	28.37 ± 4.53	21.15 ± 3.51
RBGN [40]	57.93 ± 11.56	36.95 ± 5.99	27.74 ± 5.16	20.67 ± 3.95
DSAE [27]	58.63 ± 11.23	37.50 ± 7.79	25.59 ± 5.24	20.18 ± 3.07
LGFFWM	**60.88 ± 9.63**	**39.62 ± 6.25**	**29.36 ± 3.62**	**22.56 ± 2.32**

TABLE IV
OA (%) AND SD (%) OF LGFFWM METHOD AND STATE-OF-THE-ART METHODS ON THE AID

Method	25/5	20/10	15/15	10/20
SSE [37]	46.11 ± 7.21	30.28 ± 4.90	19.94 ± 2.43	12.73 ± 1.27
DMaP [38]	43.40 ± 7.29	28.29 ± 4.78	19.38 ± 2.62	11.56 ± 1.29
SAE [39]	47.34 ± 8.42	32.12 ± 4.45	23.73 ± 3.28	13.77 ± 1.17
ZSL-LP [30]	46.77 ± 7.65	30.82 ± 4.90	21.78 ± 3.37	12.97 ± 1.06
ZSC-SA [28]	50.87 ± 8.74	33.46 ± 5.99	24.41 ± 3.83	15.89 ± 2.03
VSOP [59]	48.56 ± 7.90	32.95 ± 5.52	24.84 ± 3.04	14.03 ± 2.47
f-CLSWGAN [60]	50.68 ± 11.25	33.89 ± 5.72	24.95 ± 2.96	17.26 ± 3.06
CYCLEWGAN [61]	52.37 ± 10.47	35.94 ± 5.46	25.28 ± 2.66	17.89 ± 2.86
RBGN [40]	51.99 ± 11.32	36.27 ± 5.65	24.83 ± 3.07	16.83 ± 3.14
DSAE [27]	53.49 ± 8.58	35.32 ± 5.17	25.92 ± 3.92	17.65 ± 2.52
LGFFWM	**56.96 ± 8.14**	**37.42 ± 4.96**	**26.55 ± 3.88**	**18.36 ± 3.35**

TABLE V
OA (%) AND SD (%) OF LGFFWM METHOD AND STATE-OF-THE-ART METHODS ON THE NWPU

Method	35/10	30/15	25/20	20/25
SSE [37]	33.36 ± 3.58	23.30 ± 2.48	16.88 ± 2.29	12.94 ± 1.46
DMaP [38]	49.53 ± 6.31	38.07 ± 4.83	28.15 ± 3.86	23.95 ± 2.60
SAE [39]	44.81 ± 4.73	35.07 ± 3.91	24.65 ± 3.71	20.77 ± 2.02
ZSL-LP [30]	47.00 ± 6.64	36.45 ± 4.58	26.71 ± 3.43	22.90 ± 2.47
ZSC-SA [28]	48.40 ± 6.36	37.55 ± 4.54	28.27 ± 3.47	23.69 ± 2.38
VSOP [59]	45.32 ± 5.71	36.09 ± 4.63	25.44 ± 3.13	22.18 ± 2.00
f-CLSWGAN [60]	45.35 ± 6.37	38.97 ± 4.93	30.06 ± 2.96	24.31 ± 2.57
CYCLEWGAN [61]	46.87 ± 5.99	39.85 ± 4.71	31.17 ± 2.66	25.06 ± 2.74
RBGN [40]	44.68 ± 6.14	40.31 ± 4.89	31.91 ± 3.07	24.89 ± 2.44
DSAE [27]	51.52 ± 6.91	**41.94 ± 4.61**	31.85 ± 3.32	25.20 ± 2.17
LGFFWM	**51.96 ± 5.88**	40.86 ± 4.72	**32.18 ± 2.96**	**26.42 ± 2.56**

As shown in Tables III–V, LGFFWM achieves OA values of 60.88%, 56.96%, and 51.96% on the three datasets, which is significantly superior to other comparative methods. Compared to the DSAE method, LGFFWM improves OA by an average of 2.25%, 3.47%, and 0.44% and reduces SD by an average of 1.60%, 0.44%, and 1.03%. LGFFWM only exhibits slightly lower accuracy when the ratio of seen/unseen classes is low (Table V, ratio of 30/15). As a zero-shot RSSC method, DSAE imposes a discriminative distance metric constraint to enhance the discriminative ability of the semantic auto-encoder but overlooks the construction of a more discriminative visual space. Regarding the two zero-shot RSSC, SSE, and SAE, both have much lower OA compared to DSAE and LGFFWM. This also indicates significant differences between typical natural scenes and remote sensing scenes in terms of background complexity, land feature composition, and other factors. Hence, traditional ZSL methods are not directly applicable to RSSC tasks. The other two zero-shot RSSC methods are ZSL-LP and ZSC-SA.

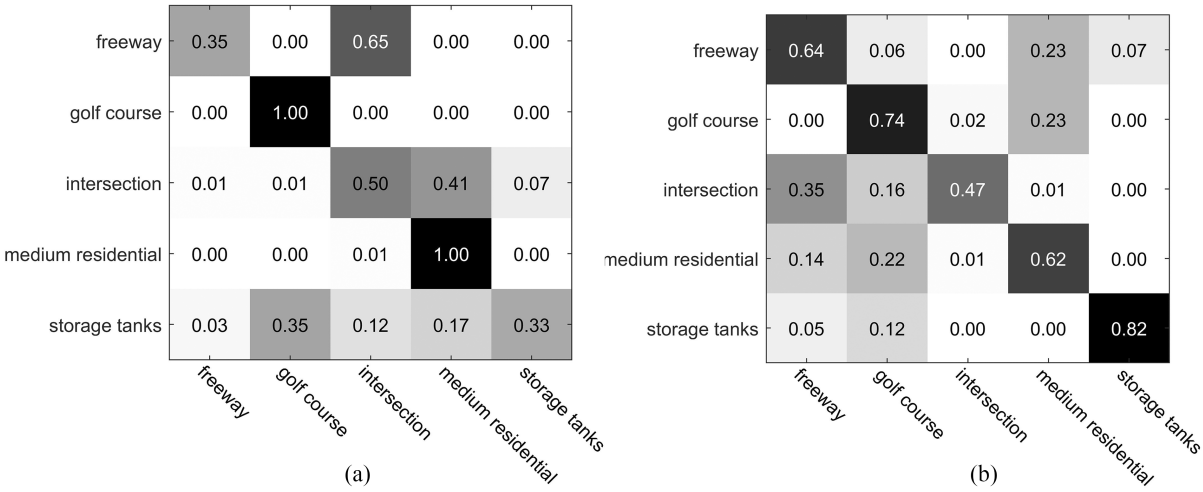


Fig. 4. CM of the two methods on the UCM. (a) DSAE, the AA is 63.6%; (b) OUR, the AA is 65.80%.

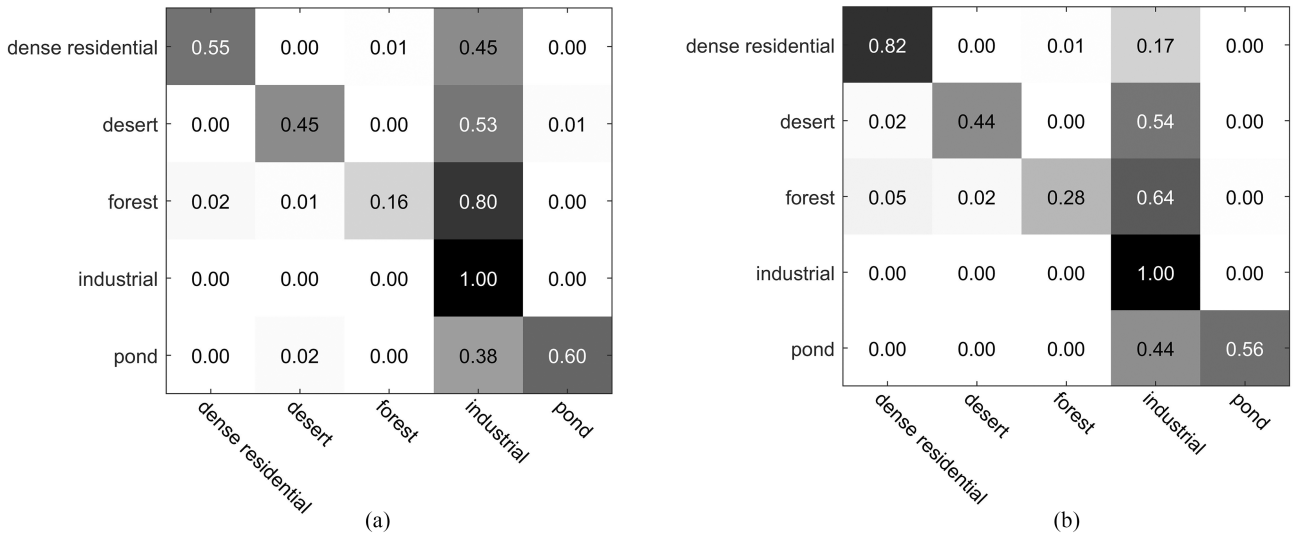


Fig. 5. CM of the two methods on the AID. (a) DSAE, the AA is 64.8%; (b) OUR, the AA is 65.70%.

Where, ZSL-LP has constructed a semantic directed graph in the semantic space to describe the relationships between seen and unseen classes, and ZSC-SA employs a semisupervised Sammon embedding algorithm to make the semantic space and visual space prototypes more consistent in class structure. Their highest accuracies are only 49.01% and 50.87%, Consequently, LGFFWM demonstrates enhanced competitiveness. Our approach, akin to ZSL-LP [30], establishes a connection between seen and unseen classes. However, ZSL-LP [30], ZSC-SA [28], and DASE [27] overlook distinctive local landscape features. This comparative analysis facilitates a comprehensive and objective assessment of LGFFWM. The experimental outcomes affirm the efficacy and dependability of the LGFFWM.

Furthermore, we have used CM to analyze the recognition performance of LGFFWM and the second-best performing DSAE for each unseen class and evaluated the AA based on this analysis, as shown in Figs. 4–6.

As shown in Fig. 4, for UCM with a ratio of seen/unseen classes set at 16/5, the unseen classes include "freeway," "golf course," "intersection," "medium residential," and "storage tanks." We observed that compared to DSAE, LGFFWM improved the detection accuracy by 29% and 49% in the "freeway," and "storage tanks" categories, respectively, compared to DSAE. Despite declines in other categories, AA improved from 63.6% to 65.80%.

As shown in Fig. 5, for AID with a ratio of seen/unseen classes set at 25/5, the unseen classes include "dense residential," "desert," "forest," "industrial," and "pond." We observed that compared to DSAE, LGFFWM improved the detection accuracy of the two classes "dense residential" and "forest" by 27% and 12%, respectively. Although there are some decreases in accuracy for other classes, overall AA has increased from 64.8% to 65.70%. As shown in Fig. 6, for NWPU with a ratio of seen/unseen classes set at 35/10, the unseen classes include "airport," "basketball court," "circular farmland," "cloud," "dense residential," "desert," "harbor," "intersection," "medium residential," and "sparse residential." We observed that compared to the DSAE method, LGFFWM improved the detection

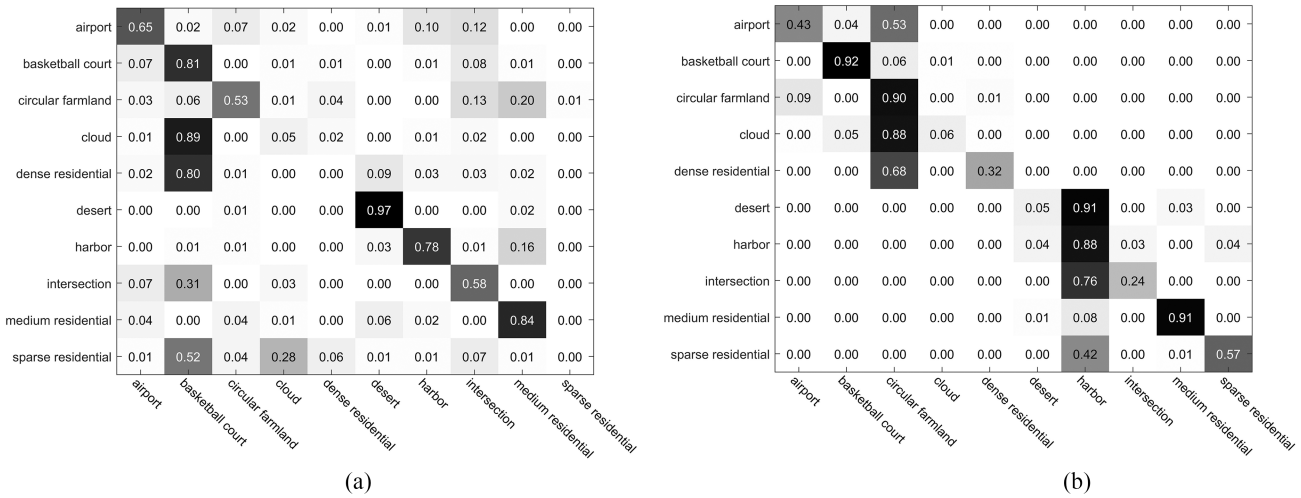


Fig. 6. CM of the two methods on the NWPU. (a) DSAE, the AA is 52.1% and (b) OUR, the AA is 52.9%.

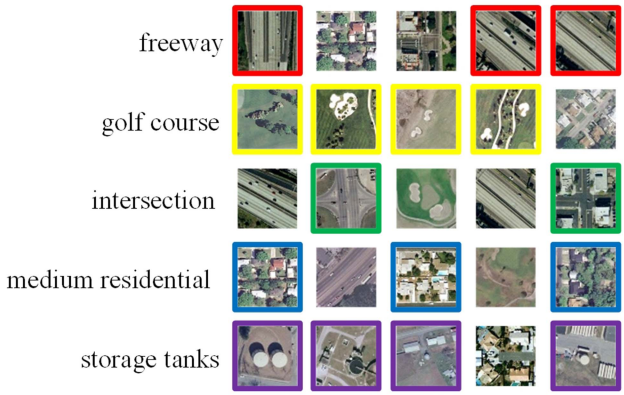


Fig. 7. Classification visualization for the UCM, where red, yellow, green, blue, and purple boxes denote "freeway," "golf course," "intersection," "medium residential," and "storage tanks," respectively.

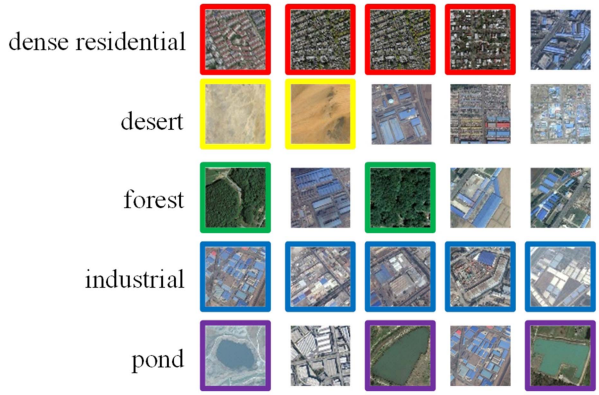


Fig. 8. Classification visualization for the AID, where red, yellow, green, blue, and purple boxes denote "dense residential," "desert," "forest," "industrial," and "pond," respectively.

accuracy of the "basketball court," "circular farmland," "cloud," "dense residential," "harbor," "medium residential," and "sparse residential" categories by 11%, 37%, 1%, 32%, 10%, 7%, and 57%, respectively. Although there are some decreases in accuracy for other classes, overall AA has increased from 52.1% to 52.9%.

To further illustrate which unseen class scenes are misclassified into which other unseen classes, we have visualized the classification results for each unseen class sample, as shown in Figs. 7–9.

Firstly, for UCM, we have used red, yellow, green, blue, and purple boxes to represent the unseen classes "freeway," "golf course," "intersection," "medium residential," and "storage tanks," respectively, as shown in Fig. 7. Three scenes of "freeway," and "medium residential" were correctly classified; Four scenes of "golf course," and "storage tanks" were correctly classified; Only two scenes of the "intersection" were correctly classified, two of which were classified as "freeway," and one as "golf course". The visualization results are consistent with the corresponding CM.

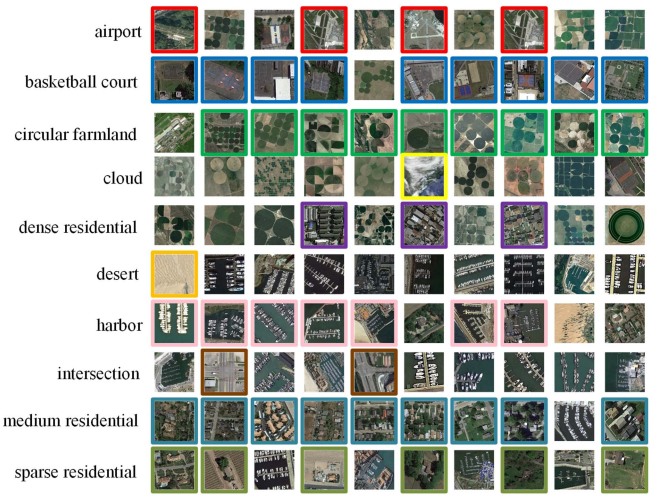


Fig. 9. Classification visualization for the NWPU, where red, blue, green, yellow, purple, orange, pink, brown, aqua, and olive boxes denote "airport," "basketball court," "circular farmland," "cloud," "dense residential," "desert," "harbor," " intersection," "medium residential," and "sparse residential," respectively.

TABLE VI
COMPARISON OF MODEL PERFORMANCE BASED ON THE UCM AND AID DATASETS

Architecture	Parameters (M)	UCM (16/5)		AID (25/5)	
		Time Cost (s/epoch)	OA (%)	Time Cost (s/epoch)	OA (%)
VGGNet	143	24.78	43.90	86.01	38.62
LGFFWM	287	456.37	**60.88**	715.37	**59.96**

Next, for AID, we have used red, yellow, green, blue, and purple boxes to represent the unseen classes "dense residential," "desert," "forest," "industrial," and "pond," respectively, as shown in Fig. 8. One scene in the "dense residential," and scene categories is classified as "industrial." Two scenes in each of "desert," and "forest" is correctly classified, and the other three are classified as "industrial." All five scenes in "industrial" are correctly classified; three of the scenes in "pond" are correctly classified. Through the visual results, we found that in the misclassifications, a majority were classified as "industrial," which prompted us to further investigate the relationships between the classes to achieve ZSL.

Finally, in the NWPU dataset, we have used red, blue, green, yellow, purple, orange, pink, brown, aqua, and olive boxes to represent "airport," "basketball court," "circular farmland," "cloud," "dense residential," "desert," "harbor," " intersection," "medium residential," and "sparse residential," respectively. The results are shown in Fig. 9. Overall, the visualization results are basically consistent with the corresponding CM.

3) Time Efficiency: To assess the temporal efficiency of LGFFWM, we measured the training time and the count of weight parameters per epoch for evaluation [63] and compared it against VGGNet. We have conducted experiments using a split ratio of 16/5 for the UCM dataset and 25/5 for the AID dataset, as described in Table VI.

From Table VI, we observe that LGFFWM takes nearly ten times as long as VGGNet, but OA improves by 16.98% and 21.34% on the UCM and AID datasets, respectively. This is because we used VGGNet's fully connected layer output as global features, as well as the softmax layer to output local features of the local landscape, resulting in a change in the number of network layers. Analysis of the table reveals that LGFFWM exhibits a double increase in parameters compared to VGGNet, yet attains superior performance. In summary, LGFFWM can achieve better OA within the allowable range of time loss, achieving a balance between accuracy and time efficiency.

IV. DISCUSSION

A. Analysis of the Effectiveness of LGFF

To demonstrate the effectiveness of LGFF, we have conducted experiments based on the UCM, AID, and NWPU, where the ratios of seen/unseen classes set at 16/5, 25/5, and 35/10, respectively (the same settings used in Sections IV-B, C, and D), as shown in Table VII.

It can be seen in Table VII that the experiments on the three datasets yielded the same conclusion: Compared to extracting only local or global features, using the LGFF designed in this article significantly improves OA. The use of local features alone leads to a noticeable decrease in OA compare to using global features alone, with decreases of 14.30%, 10.76%, and 12.29%, respectively. This indicates that global features are more discriminative. The kappa coefficients reached 0.5791, 0.5264, and 0.4079, respectively. This fully demonstrates that in the task of zero-shot RSSC, it is feasible and necessary to introduce local landscape features on the basis of global features, as they complement each other.

TABLE VII
EFFECTIVENESS ANALYSIS OF LGFF

Dataset	Visual feature	OA (%)	Kappa
UCM	F_Local	29.60	0.1205
	F_Global	43.90	0.3023
	F_Fusion	**60.88**	**0.5791**
AID	F_Local	27.86	0.0592
	F_Global	38.62	0.2010
	F_Fusion	**56.96**	**0.5264**
NWPU	F_Local	32.36	0.1544
	F_Global	44.65	0.3089
	F_Fusion	**51.96**	**0.4079**

Additionally, the widely used visualization technique, Gradient-weighted Class Activation Mapping (Grad-CAM) [64], is employed for further efficacy analysis of LGFF, as depicted in Fig. 10.

Among them, brighter regions in the feature map have higher discrimination. In contrast, ResNet50 (Residual Network 50) only focuses on a limited portion of the scene for prediction, while LGFF can capture details that represent semantic features in images with complex backgrounds, covering almost the entire target region, and has strong performance in remote sensing scene datasets. Specifically visible, LGFF can start from the overall content of the scene and focus on multiple local objects related to the scene category in terms of global features. It can be inferred that LGFF has stronger feature extraction ability and can effectively learn discriminative features of the target region.

B. Analysis of the Effectiveness of WM Loss

To further analyze the effectiveness of WM Loss, we have conducted experiments by replacing the loss functions in LGFFWM with classic loss functions, including Cross-Entropy Loss (CE Loss), Mean Squared Error Loss (MSE Loss), and Least Square Embedding Loss (LSE Loss), as shown in Table VIII.

It can be seen in Table VIII that the experiments on the three datasets yielded the same conclusion: using the WM Loss

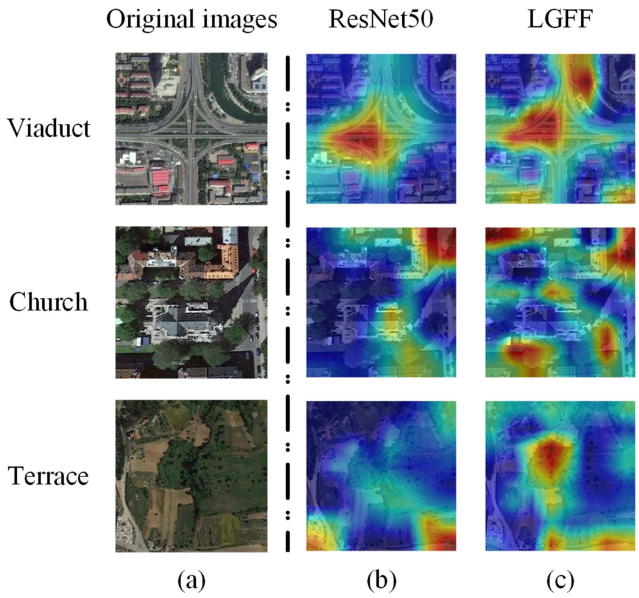


Fig. 10. Visualization results of Grad-CAM. (a) Original images. (b) ResNet50. (c) LGFF.

TABLE VIII
EFFECTIVENESS ANALYSIS OF WM LOSS

Dataset	Loss	OA (%)	Kappa
UCM	CE Loss	38.80	0.2392
	MSE Loss	43.62	0.2995
	LSE Loss	48.24	0.3550
	WM Loss	**60.88**	**0.5791**
AID	CE Loss	32.88	0.1262
	MSE Loss	46.21	0.2912
	LSE Loss	46.83	0.2849
	WM Loss	**56.96**	**0.5264**
NWPU	CE Loss	28.94	0.1123
	MSE Loss	37.34	0.2174
	LSE Loss	43.88	0.2992
	WM Loss	**51.96**	**0.4079**

designed in this article significantly improves OA compared to classic CE Loss, MSE Loss, and LSE Loss; using the designed WM Loss results in a substantial increase in OA compared to classic CE Loss, MSE Loss, and LSE Loss, with the highest improvements being 27.08%, 17.26%, and 13.13%, respectively. This demonstrates the effectiveness and feasibility of the WM Loss designed in this article.

C. Analysis of the Impact of Semantic Vector Dimension

While constructing the semantic space, there are three different dimensions for the semantic vectors output by the Word2Vec model pretrained on Wikipedia: 100, 300, and 500. Therefore, we further discuss the impact of different semantic vector dimensions on OA, and the results are shown in Fig. 11.

As shown Fig. 11, for vector dimensions of 100, 300, and 500, the corresponding OA values on the UCM are 0.5500, 0.6088, and 0.6020, respectively. On the AID, the corresponding OA values are 0.5342, 0.5696, and 0.5602, respectively. On the NWPU, the corresponding OA values are 0.4712, 0.5196, and 0.5048, respectively. It can be seen that the best performance is achieved when the vector dimension is 300. We believe this is because an adequate-low dimension leads to information loss and confusion, making semantic vectors overly simplified and abstract, making it difficult to distinguish subtle semantic differences, thus reducing the model's discriminative power. On the other hand, an overabundant dimension can significantly increase computational costs and may even lead to overfitting issues for the model.

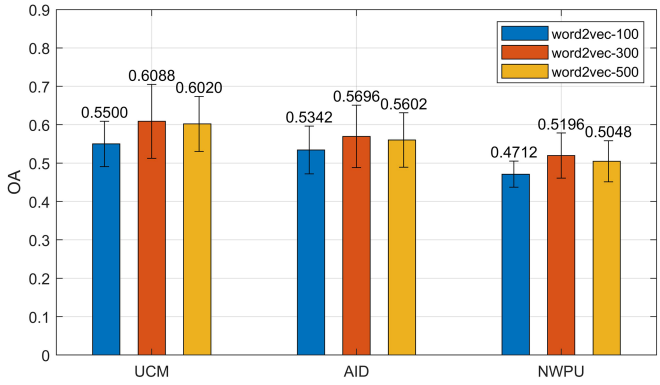


Fig. 11. OA and SD of semantic vectors with different dimensions on the UCM, AID, and NWPU.

D. Analysis of the Influence of the Image Resolution

When images are at a higher spatial resolution, more pixels are included in an image of the same object and thus, more details are found [65]. That is why this article analyzed the impact of resolution changes on OA. In the experiments, we downsampled the original images to 0.4, 0.6, and 0.8 times the resolution and tested them using LGFFWM, as shown in Fig. 12.

It can be seen that, with the increase in resolution, the three datasets show a gradual increasing trend in OA when using LGFFWM. We believe this is because higher resolution implies richer contextual information, which helps in fine-grained characterization of urban remote sensing scenes and results in a more discriminative feature space. This also indicates that changes in resolution have a significant impact on the accuracy of urban RSSC under zero-shot conditions.

E. Analysis of the Transferability of the LGFFWM

To further validate the transferability of LGFFWM, supplementary experiments are conducted on the SIRI-WHU [57] and RSSCN7 [58] datasets, using OA and kappa for assessment. Results are presented in Tables IX–X.

As shown in the Tables IX–X, the highest OA values are 53.50% and 47.37%, respectively, with little fluctuation compared to OA on the UCM, AID, and NWPU datasets. Specifically, under different seen/unseen class ratios (9/3, 7/5, 5/7, 3/9) on the SIRI-WHU dataset, the OA values are 53.50%, 39.70%, 27.42%, and 15.72%, respectively. On the RSSCN7 dataset with

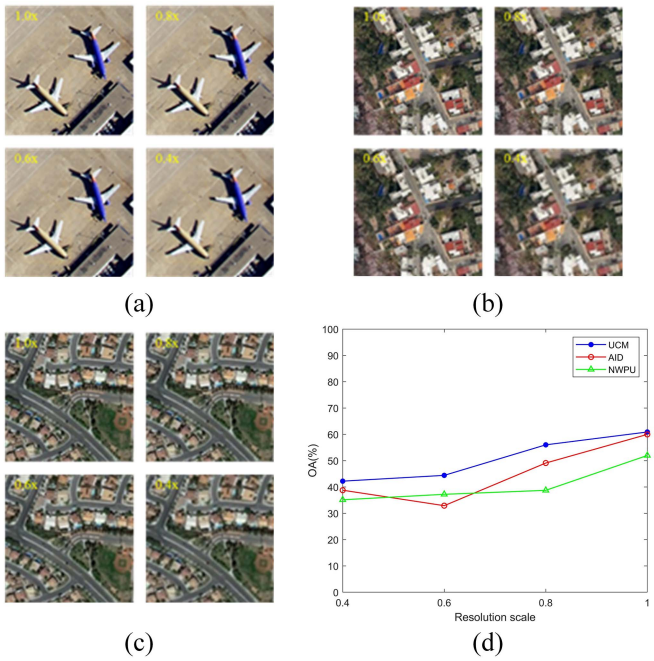


Fig. 12. Influence of the spatial resolution of images on the OA (%): (a) down-sampling results of original images at the resolution scales of 0.8, 0.6 and 0.4 based on the UCM; (b) down-sampling results of original images at the resolution scales of 0.8, 0.6 and 0.4 based on the AID; (c) down-sampling results of original images at the resolution scales of 0.8, 0.6, and 0.4 based on the NWPU; and (d) OA (%) curves of images with different resolutions based on the UCM,AID, and NWPU.

TABLE IX
CLASSIFICATION RESULTS OF SIRI-WHU DATASET WITH DIFFERENT SEEN/UNSEEN CLASS SEGMENTATION RATIOS

Seen/Unseen	OA (%)	Kappa
9/3	53.50	0.3233
7/5	39.70	0.2483
5/7	27.42	0.1547
3/9	15.72	0.0527

TABLE X
CLASSIFICATION RESULTS OF RSSCN7 DATASET WITH DIFFERENT SEEN/UNSEEN CLASS SEGMENTATION RATIOS

Seen/Unseen	OA (%)	Kappa
5/2	47.37	-0.0526
4/3	34.08	0.0125
3/4	26.43	0.0200
2/5	23.85	0.0488

different seen/unseen class ratios (5/2, 4/3, 3/4, 2/5), the OA values are 47.37%, 34.08%, 26.43%, and 23.85%, respectively. Experimental results affirm the transferability of the LGFFWM method.

V. CONCLUSION

In this article, we proposed a way out for zero-shot RSSC based on LGFF and WM Loss. Benefitting from our design LGFFWM, the model can adaptively label local landscapes in the scenes and effectively fuse them with global visual features, resulting in a more discriminative representation of urban remote sensing scenes. Building upon this, we introduced a WM Loss function based on the semantic correlation matrices to address the varying semantic relatedness between each unseen class and different seen-class scenes. This loss function can adaptively adjust the assigned weights during training based on the semantic relevance between seen-class samples and a particular unseen sample, thus compelling the model to focus more on learning from highly relevant samples. This enhances the model's discriminative ability by fully exploiting relevant semantic knowledge. We conducted extensive experiments on three datasets, achieving an OA for unseen class classification of up to 60.88%, 56.98%, and 51.96%, respectively, and SD less than 9.63%, 8.14%, and 5.88%, respectively, which significantly outperformed ten advanced baseline methods.

In future work, we plan to explore two main directions: 1) the limited number of feature types selected in the proposal frame extraction process in this article makes it difficult to construct a relatively complete feature space to reflect the particularities of the urban landscape. Therefore, how to construct a more discriminative landscape index set and evaluate its recognition ability for different urban landscapes is a research scope that we will soon carry out and 2) in real-world applications, scenes may come from both seen and unseen classes. Therefore, studying and analyzing GZSL in remote sensing scene datasets will be a direction for our future research.

REFERENCES

[1] J. Kong, Q. Sun, M. Mukherjee, and J. Lloret, "Low-rank hypergraph hashing for large-scale remote sensing image retrieval," *Remote Sens.*, vol. 12, no. 7, 2020, Art. no. 1164.
[2] F. Ye, H. Xiao, X. Zhao, M. Dong, W. Luo, and W. Min, "Remote sensing image retrieval using convolutional neural network features and weighted distance," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 10, pp. 1535–1539, Oct. 2018.
[3] L. Han, P. Li, X. Bai, C. Grecos, X. Zhang, and P. Ren, "Cohesion intensive deep hashing for remote sensing image retrieval," *Remote Sens.*, vol. 12, no. 1, 2020, Art. no. 101, doi: 10.3390/rs12010101.
[4] Y. Zhao, J. Qi, F. Korn, and X. Wang, "Scalable building height estimation from street scene images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3000718, doi: 10.1109/TGRS.2022.3206223.
[5] Y. Da, X. Gao, and M. Li, "Remote sensing image ship detection based on improved YOLOv3," in *Proc. 7th Int. Conf. Intell. Comput. Signal Process.*, 2022, pp. 1776–1781, doi: 10.1109/ICSP54964.2022.9778531.
[6] H. Li, A. Liu, X. Xie, H. Guo, H. Xiong, and X. Zheng, "Learning dense consistent features for aerial-to-ground structure-from-motion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 5089–5102, 2023.
[7] L. Sun, Y. Fang, Y. Chen, W. Huang, Z. Wu, and B. Jeon, "Multi-structure KELM with attention fusion strategy for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5539217, doi: 10.1109/TGRS.2022.3208165.
[8] A. M. Cheriyadat, "Unsupervised feature learning for aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 439–451, Jan. 2014, doi: 10.1109/TGRS.2013.2241444.
[9] J. Zhang, T. Li, X. Lu, and Z. Cheng, "Semantic classification of high-resolution remote-sensing images based on mid-level features," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 6, pp. 2343–2353, Jun. 2016.
[10] Q. Zhu et al., "Knowledge-guided land pattern depiction for urban land use mapping: A case study of Chinese cities," *Remote Sens. Environ.*, vol. 272, Apr. 2022, Art. no. 112916.

[11] Q. Zhu, Y. Sun, Q. Guan, L. Wang, and W. Lin, "A weakly pseudo-supervised decorrelated subdomain adaptation framework for cross-domain land-use classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5623913, doi: 10.1109/TGRS.2022.3170335.

[12] Z. Lv, P. Zhong, W. Wang, Z. You, J. A. Benediktsson, and C. Shi, "Novel piecewise distance based on adaptive region key-points extraction for LCCD with VHR remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5607709, doi: 10.1109/TGRS.2023.3268038.

[13] Z. Lv, P. Zhang, W. Sun, J. A. Benediktsson, J. Li, and W. Wang, "Novel adaptive region spectral–spatial features for land cover classification with high spatial resolution remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5609412.

[14] Z. Lv, H. Huang, W. Sun, M. Jia, J. A. Benediktsson, and F. Chen, "Iterative training sample augmentation for enhancing land cover change detection performance with deep learning neural network," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2023.3282935.

[15] Z. Lv et al., "Land cover change detection with heterogeneous remote sensing images: Review, progress, and perspective," *Proc. IEEE*, vol. 110, no. 12, pp. 1976–1991, Dec. 2022, doi: 10.1109/JPROC.2022.3219376.

[16] Y. Fang, Q. Ye, L. Sun, Y. Zheng, and Z. Wu, "Multiattention joint convolution feature representation with lightweight transformer for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5513814.

[17] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 3735–3756, 2020, doi: 10.1109/JSTARS.2020.3005403.

[18] O. A. B. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2015, pp. 44–51.

[19] S. Chaib, H. Liu, Y. Gu, and H. Yao, "Deep feature fusion for VHR remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 8, pp. 4775–4784, Aug. 2017.

[20] J. Gawlikowski, S. Saha, A. Kruspe, and X. X. Zhu, "An advanced Dirichlet prior network for out-of-distribution detection in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5616819, doi: 10.1109/TGRS.2022.3140324.

[21] J. Rosier, J. van Vliet, and V. Bakker, "Mapping intra-urban development trajectories in Nairobi, Kenya," in *Proc. Joint Urban Remote Sens. Event*, 2023, pp. 1–4.

[22] Y. Wang, Q. Huang, A. Zhao, H. Lv, and S. Zhuang, "Semantic network-based impervious surface extraction method for rural-urban fringe from high spatial resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4980–4998, 2021.

[23] W. Huang, Z. Yuan, A. Yang, C. Tang, and X. Luo, "TAE-net: Task-adaptive embedding network for few-shot remote sensing scene classification," *Remote Sens.*, vol. 14, no. 1, Dec. 2022, Art. no. 111, doi: 10.3390/rs14010111.

[24] B. Wang, Z. Wang, X. Sun, Q. He, H. Wang, and K. Fu, "TDNet: A novel transductive learning framework with conditional metric embedding for few-shot remote sensing image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 4591–4606, 2023, doi: 10.1109/JSTARS.2023.3263149.

[25] Q. Zeng, J. Geng, K. Huang, W. Jiang, and J. Guo, "Prototype calibration with feature generation for few-shot remote sensing image scene classification," *Remote Sens.*, vol. 13, no. 14, Jul. 2021, Art. no. 2728, doi: 10.3390/rs13142728.

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

[27] C. Wang, G. Peng, and B. De Baets, "A distance-constrained semantic autoencoder for zero-shot remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 12545–12556, 2021.

[28] J. Quan, C. Wu, H. Wang, and Z. Wang, "Structural alignment based zero-shot classification for remote sensing scenes," in *Proc. IEEE Int. Conf. Electron. Commun. Eng.*, 2018, pp. 17–21.

[29] C. Wu, Y. Wei, H. Wang, Y. Liu, S. Li, and J. Quan, "Transductive zero-shot classification algorithm for remote sensing image scenes," *Application Res. Comput.*, vol. 37, no. 5, pp. 1597–1600, May 2022.

[30] A. Li, Z. Lu, L. Wang, T. Xiang, and J.-R. Wen, "Zero-shot scene classification for high spatial resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 4157–4167, Jul. 2017, doi: 10.1109/TGRS.2017.2689071.

[31] Z. Li, D. Zhang, Y. Wang, D. Lin, and J. Zhang, "Generative adversarial networks for zero-shot remote sensing scene classification," *Appl. Sci.*, vol. 12, no. 8, 2022, Art. no. 3760.

[32] S. Ma, C. Liu, Z. Li, and W. Yang, " Integrating adversarial generative network with variational autoencoders towards cross-modal alignment for zero-shot remote sensing image scene classification," *Remote Sens.*, vol. 14, no. 18, 2022, Art. no. 4533.

[33] Q. Zhu, Y. Zhong, B. Zhao, G.-S. Xia, and L. Zhang, "Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 6, pp. 747–751, Jun. 2016.

[34] J. Ma et al., "Remote sensing scene classification based on global and local consistent network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 537–540, doi: 10.1109/IGARSS39084.2020.9323281.

[35] F. Li and J. Wang, "Remote sensing image scene classification via regional growth-based key area fine location and multilayer feature fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 5500905, doi: 10.1109/LGRS.2022.3233374.

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

[37] Z. Zhang and V. Saligrama, "Zero-shot learning via semantic similarity embedding," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4166–4174, doi: 10.1109/ICCV.2015.474.

[38] Y. Li, D. Wang, H. Hu, Y. Lin, and Y. Zhuang, "Zero-shot recognition using dual visual-semantic mapping paths," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5207–5215, doi: 10.1109/CVPR.2017.553.

[39] E. Kodirov, T. Xiang, and S. Gong, "Semantic autoencoder for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4447–4456.

[40] Y. Xing, S. Huang, L. Huangfu, F. Chen, and Y. Ge, "Robust bidirectional generative network for generalized zero-shot learning," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2020, pp. 1–6.

[41] L. Zhang, T. Xiang, and S. Gong, "Learning a deep embedding model for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3010–3019.

[42] X. Cheng, J. Lu, J. Feng, B. Yuan, and J. Zhou, "Scene recognition with objectness," *Pattern Recognit.*, vol. 74, pp. 474–487, 2018.

[43] Z.-R. Huang, "Fusion of complex networks-based global and local features for feature representation," in *Proc. Int. Conf. Mach. Learn. Cybern.*, 2021, pp. 1–6.

[44] Q. Zhao, B. Wang, B. Zhou, J. Di, and L. Chen, "Remote sensing image surface feature classification based on VGG-UNet," in *Proc. Int. Conf. Comput. Inf. Sci. Artif. Intell.*, 2021, pp. 1043–1047.

[45] P. B, K. Maharajan, and R. Srikanteswara, "Deep learning based image classification using small VGG net architecture," in *Proc. IEEE 2nd Mysore Sub Sect. Int. Conf.*, 2022, pp. 1–6.

[46] F. Li and J. Wang, "Remote sensing image scene classification via regional growth-based key area fine location and multilayer feature fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 5500905, doi: 10.1109/LGRS.2022.3233374.

[47] J. Chen, J. Yi, A. Chen, and Z. Jin, "EFCOMFF-Net: A multiscale feature fusion architecture with enhanced feature correlation for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5604917.

[48] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, " Efficient estimation of word representations in vector space," in *Proc. Int. Conf. Learn. Representations*, 2013, pp. 1–12.

[49] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proc. 18th Int. Conf. Pattern Recognit.*, 2006, vol. 3, pp. 850–855.

[50] X. Huang and L. Zhang, " Multidirectional and multiscale morphological index for automatic building extraction from multispectral GeoEye-1 imagery," *Photogrammetric Eng. Remote Sens.*, vol. 77, pp. 721–732, 2011.

[51] Y. Fu et al., "Winter wheat nitrogen status estimation using UAV-based RGB imagery and Gaussian processes regression," *Remote Sens.*, vol. 12, no. 22, 2020, Art. no. 3778.

[52] C. Deng and C. Wu, " BCI: A biophysical composition index for remote sensing of urban environments," *Remote Sens. Environ.*, vol. 127, pp. 247–259, 2012.
[53] R. Tao and J. Qiao, "Fast component tree computation for images of limited levels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3059–3071, Mar. 2023.
[54] Y. Yang and S. Newsam, " Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. ACM SIGSPATIAL Int. Workshop Adv. Geographic Inf. Syst.*, no. 10, 2010, pp. 270–279.
[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, doi: 10.1109/TGRS.2017.2685945.
[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] B. Zhao, Y. Zhong, G.-S. Xia, and L. Zhang, "Dirichlet-derived multiple topic scene classification model for high spatial resolution remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2108–2123, Apr. 2016, doi: 10.1109/TGRS.2015.2496185.
[58] Q. Zou, L. Ni, T. Zhang, and Q. Wang, "Deep learning based feature selection for remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015, doi: 10.1109/LGRS.2015.2475299.
[59] H. Wu, Y. Yan, S. Chen, X. Huang, Q. Wu, and M. K. Ng, "Joint visual and semantic optimization for zero-shot learning," *Knowl.-Based Syst.*, vol. 215, 2021.
[60] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5542–5551.
[61] R. Felix, B. G. Vijay Kumar, and G. Carneiro, " Multi-modal cycle-consistent generalized zero-shot learning," in *Proc. Eur. Conf. Comput. Vis.*, 2018.
[62] X. Wang, L. Duan, C. Ning, and H. Zhou, "Relation-attention networks for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 422–439, 2022, doi: 10.1109/JSTARS.2021.3135566.
[63] X. Chen, Z. Han, Y. Li, M. Ma, S. Mei, and W. Cheng, "Attention-aware deep feature embedding for remote sensing image scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1171–1184, 2023, doi: 10.1109/JSTARS.2022.3229729.
[64] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.
[65] C. Wang et al., "Earthquake-damaged buildings detection in very high-resolution remote sensing images based on object context and boundary enhanced loss," *Remote Sens.*, vol. 13, 2021.

Chao Wang received the B.S. degree in information engineering and the M.S. degree in communication and information system from the China University of Mining and Technology, Xuzhou, China, in 2007 and 2010, respectively, and the Ph.D. degree in computer application technology from Hohai University, Nanjing, China, in 2014.

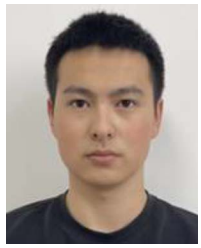
He is currently an Associate Professor with the School of Electronics and Information Engineering, Nanjing University of Information Science and Technology, Nanjing. His current research interests include remote sensing image processing and applications, machine learning, deep learning, and pattern recognition.

Junyong Li received the B.S. degree in electronic information engineering from the Henan Institute of Technology, Xinxiang, China, in 2021. He is currently working toward the M.S. degree in electronic information with the Nanjing University of Information Science and Technology, Nanjing, China.

His research interests include remote sensing image processing and deep learning.

Ahmed Tanvir received the B.S. degree in electronic information engineering, in 2023, from the Nanjing University of Information Science and Technology, Nanjing, China, where he is currently working toward the M.S. degree in information and communication engineering.

His research interests include the intersection of remote sensing and deep learning.



Jiajun Yang received the B.S. degree in electronic information engineering from the Nanyang Institute of Technology, Nanyang, China, in 2022. He is currently working toward the M.S. degree in electronic information with the Nanjing University of Information Science and Technology, Nanjing, China.

His research interests include zero-shot learning and remote sensing scene classification.

Tao Xie received the M.S. degree in materials from the Guizhou University of Technology, Guiyang, China, in 2002, and the Ph.D. degree in electromagnetic fields and microwave technology from Shanghai Jiao Tong University, Shanghai, China, in 2005.

He is currently a Professor with the School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing. His current research interests include electromagnetic fields and microwave technology, satellite oceanography, satellite meteorology, and polar remote sensing.

Liqiang Ji received the B.S. degree in electronic information engineering from Lishui University, Lishui, China, in 2022. He is currently working toward the M.S. degree in electronic information with the Nanjing University of Information Science and Technology, Nanjing, China.

His research interests include machine learning, deep learning, and remote sensing image processing.

Tong Zhang received the B.S. degree in communication engineering from Zhonghuan Information College, Tianjin University of Technology, Tianjin, China, in 2022. She is currently working toward the M.S. degree in electronic information with the Nanjing University of Information Science and Technology, Nanjing, China.

Her research interests include image enhancement and position measurement.