Remote-Sensing Scene Classification via Multistage Self-Guided Separation Network

Junjie Wang, Wei Li, Mengmeng Zhang, Ran Tao, Jocelyn Chanussot

Remote Sensing Scene Classification via Multi-Stage Self-Guided Separation Network

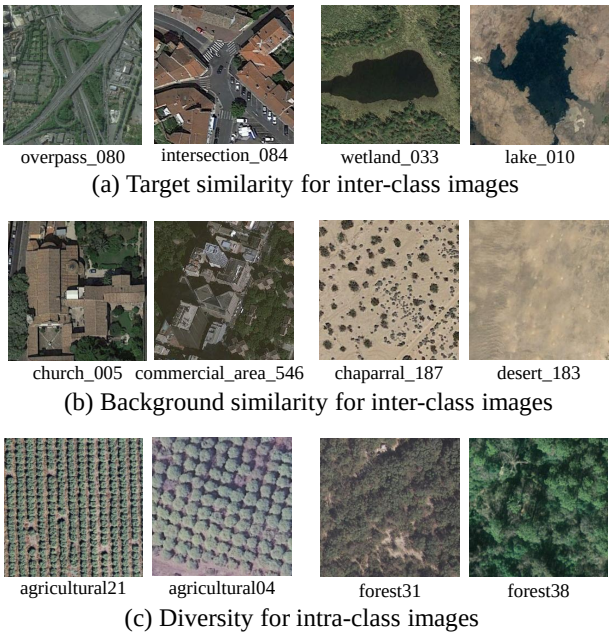
Junjie Wang, Wei Li, *Senior Member, IEEE*, Mengmeng Zhang, Ran Tao, *Senior Member, IEEE*, Jocelyn Chanussot, *Fellow, IEEE*

***Abstract*—In recent years, remote sensing scene classification is one of research hotspots and has played an important role in the field of intelligent interpretation of remote sensing data. However, various complex objects and backgrounds form a variety of remote sensing scenes through spatial combination and correlation, which brings great challenges to accurately classify different scenes. Among them, the insufficient feature difference brought about the unbalanced change of background and target between inter-class sample and the feature representation inconsistency caused by the difference of representation among the intra-class samples have become obstacles to effectively distinguish different scene images. To address these issues, a Multi-stage Self-Guided Separation Network (MGSNet) is proposed for remote sensing scene classification. First of all, different from the previous work, it attempts to utilize the background information outside the effective target in the image as a decision aid through a target-background separation strategy to improve the distinguish ability between target similarity-background difference samples. In addition, the diversity of feature concerns among different network branches is expanded through contrastive regularization to improve the separation of target-background information. Additionally, a self-guided network is proposed to find common features between intra-class samples and improve the consistency of feature representation. It combines the texture and morphological features of images to guide feature learning, effectively reducing the impact of intra-class differences. Extensive experimental results on three benchmark demonstrate that MGSNet can achieve better classification performance compared to the state-of-the-art approaches.**

***Index Terms*—Remote sensing, scene classification, target-background separation strategy, self-guided network.**

I. Introduction

REMOTE sensing scene classification is a research hotspot in the intelligent interpretation task of remote sensing data, which aims to focus on high-level semantic information in scene images and classify them into corresponding scene categories, while providing scene-level data understanding and decision aids for many practical applications, such as land use and cover monitoring, urban development and planning, and natural disaster response [1]–[5]. In recent years, with the continuous progress of remote sensing observation technology, the application scope of remote sensing images has become more and more extensive, how to make full use of the increasing number of remote sensing images for intelligent earth observation becomes extra important [6]–[9]. Therefore, a great deal of research has been performed in the last decades to achieve scene classification [10] [11]. Feature extraction plays an important role in the scene classification task, and the expressiveness of the extracted features also determines the performance of the classification. According to the types of extracted features, existing remote sensing scene classification work can be roughly divided into two directions: handcrafted feature-based methods, and feature learning-based methods [12] [13].



(a) Target similarity for inter-class images

(b) Background similarity for inter-class images

(c) Diversity for intra-class images

Fig. 1. The similarity of inter-class images and the diversity of intra-class images.

Early methods for scene classification are mainly based on handcrafted feature extraction. These methods take the professional knowledge of practitioners as a priori to extract features such as color, texture, shape, and spatial and spectral information from scene images for decision support of scene classification. Literature [14] proposed an object-oriented classification method, which combined improved color structure code (CSC) to classify high-resolution data. In [15], a global morphological texture descriptor was designed to explore the

This paper is supported by the National Key R&D Program of China (Grant No. 2021YFB3900502). (Corresponding author: Wei. Li)

J. Wang, W. Li, M. Zhang, and R. Tao are with the School of Information and Electronics, Beijing Institute of Technology, and Beijing Kay Laboratory of Fractional Signals and Systems, Beijing 100081, China (e-mail: junjiewang@bit.edu.cn, liwei089@ieee.org, mengmengzhang@bit.edu.cn, rantao@bit.edu.cn).

Jocelyn Chanussot is with the LJK, CNRS, INRIA, Grenoble INP, University of Grenoble Alpes, 38000 Grenoble, France, and also with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: jocelyn.chanussot@gipsa-lab.grenoble-inp.fr).

potential of multi-scale texture descriptors in scene classification tasks. In [16], a unique invariant feature extraction method was proposed, these features exhibited invariance to image scale and rotation, and can robustly identify objects in clutter and occlusion. Reference [17] studied the importance of directional gradient histograms in effective target detection. However, it is worth noting that methods relying on handcrafted-feature have good results on images with the same texture structure or spatial distribution, but they are still limited when faced with images with complex scenes, due to the fact that artificially designed features can significantly affect the representational ability of image features. In order to make up for the limitations of handcrafted feature-based methods, automatic feature extraction from images has gradually become a new feature extraction method.

This has been followed by the rise of feature learning-based methods. Feature learning-based methods [18] [19] are able to learn the corresponding adaptive functions from the original pixel values or handcrafted features through sparse coding, autoencoder and other means, so as to obtain a more appropriate image representation for scene classification [12]. Compared to methods based on handcrafted-feature, the difference lies in the automatic learning of relevant features, rather than relying on manually designed features for discrimination. However, since the above mentioned methods do not utilize the corresponding label information, their feature extraction and learning capability are limited, and it is not conducive to further improvement of scene classification performance.

In recent years, with the development of artificial intelligence, deep learning methods provide excellent performance, and have made important progress in image classification, object recognition, and semantic segmentation [20]–[23]. The method based on deep feature learning enables the model self-learning more powerful, abstract, and meaningful features through the deep network architecture, avoiding the defects of manually designed features. At the same time, the introduction of label information enables the model to learn more accurate distinctions between classes and improving classification performance [24] [25] [26]. In literature [27], a multi-instance densely connected ConvNet was proposed, which treated scene classification as a multi-instance learning problem to further investigate local semantics. In [28], a new discriminant function was introduced to improve the training effect of the model. For this reason, in addition to the common classification loss, the metric learning regularization term was also introduced and applied to CNN features, making the model more distinctive. Liu et al. [29] constructed a two-branch multiscale convolutional neural network using a fixed-scale network and a variable-scale network to cope with the scale variation of the target in the image, in addition to adding a similarity measurement layer to ensure the similarity between the original image and the scaled image features. In [30], an enhanced attention module was designed to enhance the feature extraction and generalization capabilities of the deep neural networks to better recognize small objects in scenes with complex backgrounds. Reference [31] attempted to use a pre-trained convolutional neural network model to reduce training time, and proposed a fusion strategy to integrate the multi-layer features of the model. In [32], the convolutional local attention module was embedded in all down-sampling and residual blocks of the ResNet backbone to construct a multi-branch local attention network, which placed the convolutional channel attention module and the local spatial attention module in parallel to obtained channel and spatial attention, respectively, which helped to emphasize the main targets in complex backgrounds and mentions the representational power of features.

Although the above works have tried to solve the problems in scene classification tasks from various perspectives, there are still some problems when faced with the classification problems in actual complex scenes. 1) Feature confusion due to the difference of target-background imbalance changes. The similarity between backgrounds and targets in inter-class images leads to easily confused and poorly discriminative extracted features, where the well-known issue of target similarity between inter-class images hinders the effective discrimination of different scenes. For example, the similarity of the targets of the samples shown in Fig.1 (a) leads to a high degree of confusion between different scenes. Furthermore, scene labels are usually determined by internal key targets in previous methods, ignoring the decision supplement that background information can provide. For example, the difference in background information in Fig.1 (a) can provide more information supplements for distinguishing classes with similar targets. 2) Inconsistent feature representation due to the diversity for intra-class sample. As shown in Fig.1 (c), in the case of non-uniform sample collection and a large number of disturbances, such as seasonal changes, lighting intensity, and regional variations, the representation of intra-class samples show diversity, which poses challenges to the robustness of the model.

To solve the above problems, a novel multi-stage self-guided separation network (MGSNet) is proposed for remote sensing scene classification. First, in order to solve the problem of insufficient feature representation caused by the similarity due to the unbalanced change of target-background, attempts have been done to separate the background and target to extract corresponding features for the first time in the field of remote sensing scene classification. Specifically, the target and background are distinguished by binary segmentation, followed by feature extraction of the target and background separately using a two-branch network. In addition, the differences between target features and background features are enlarged by introducing contrastive regularization [33], so that the key target features are fully mined while retaining the background information, providing assistance for better distinguishing scenes with similar targets. Secondly, to improve the feature consistency among intra-class samples with different representation, a self-guided network is developed, which utilizes the texture features and image morphological features of the samples to guide the learning of the main branch to reduce the impact of intra-class differences. Compared with previous methods, MGSNet has the advantage of not blindly processing the entire image in a uniform manner, but using different feature extraction branches to simultaneously extract the target, background, and overall features, which is

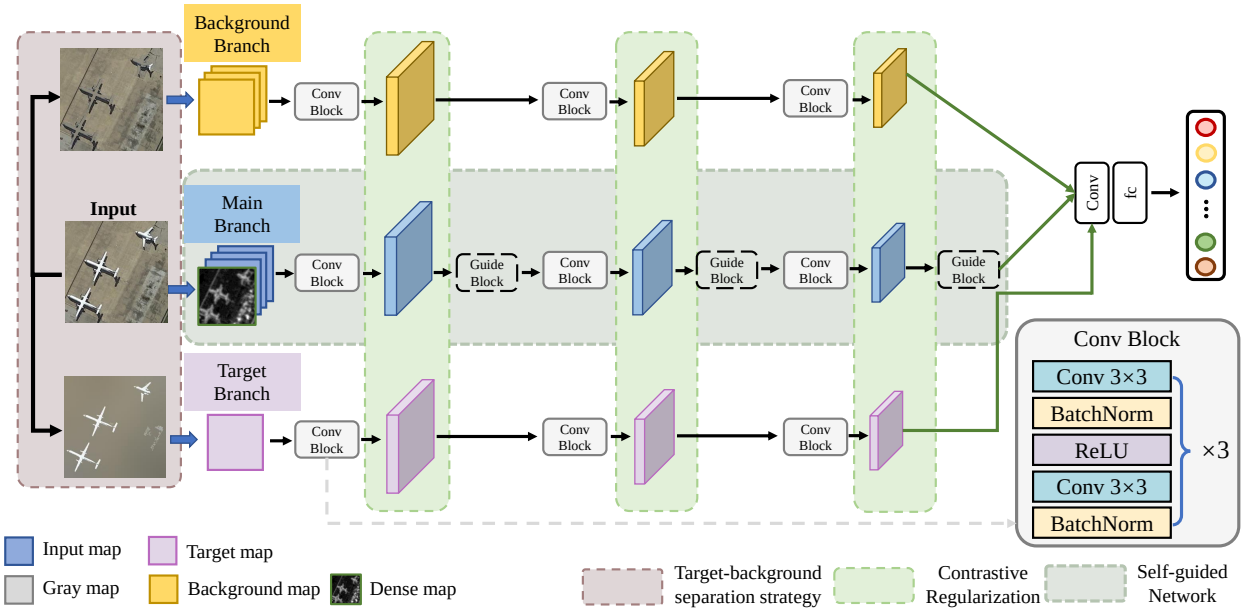


Fig. 2. Schematic illustration of the proposed MGSNet. First, for input images, a target-background separation strategy is applied to extract the target and background information in the scene and send them into the corresponding feature extraction branch. Next, the contrast regularization method is utilized to add additional constraints to each branch, thus increasing the separation of target-background information. Furthermore, the self-guided network is proposed to guide the learning of main branch feature by exploiting texture and topological information in images.

not considered by previous methods. In addition, a guidance mechanism based on density and structural information is designed for the feature extraction of the main branch, so that the network model can better focus on key areas, improving the discriminability and understanding of the model.

In summary, the main contributions are summarized as follows.

1) Aiming at mitigating the impact of background-target imbalance changes and improving the discrimination between similar samples, innovatively separating the target from the background and extracting relevant features in the remote sensing scene classification task, which is a way that has not been involved in previous work. To this end, a target-background separation strategy is designed to reduce the confusion between similar targets by extracting the key targets and important background information in the scene images separately. Meanwhile, the separation between background and target features is improved by contrastive regularization, which enables different branches to extract specific information.
2) To control the variability between intra-class samples and capture class-intrinsic features, a self-guided network is developed. Different from ordinary convolutional networks, it guides the collection and learning of features through the inherent texture and morphological features within the samples, which reduces the interference introduced in the process of sample collection and makes the feature representation between intra-class samples more unified.
3) Different from the previous methods, the proposed MGSNet not only solves the well-known problems of inter-class target similarity and intra-class image differences in scene classification tasks, but also attempts to introduce background information as a key supplement to the final classification features.

The rest of the paper is organized as follows. Section II describes the proposed MGSNet in detail, including how to separate background and target features in the scene and the architecture of the self-guided network. In section III, extensive experiments and discussions are deployed to verify the effectiveness of the proposed method and the role played by each module. Finally, section IV draws the conclusion.

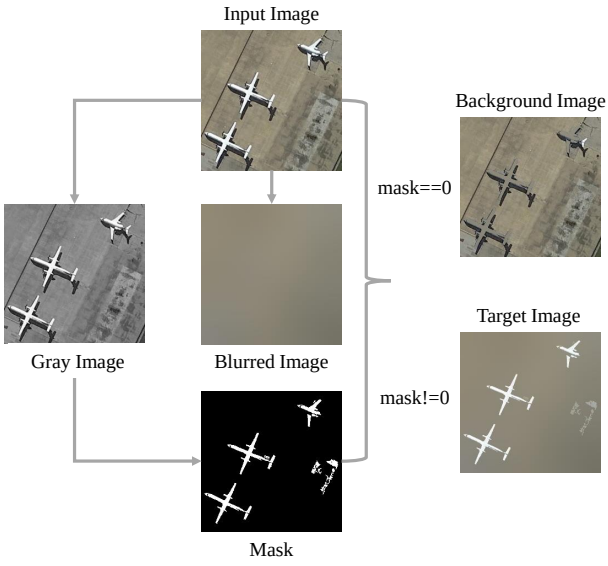


Fig. 3. The illustration of target-background separation strategy.

II. Proposed MGSNet Framework

The framework of the proposed MGSNet is illustrated in Fig. 2. First, the input image is separated into background and target, after which it is sent to the corresponding branches to extract meaningful features. In the gap of feature extraction in multiple stages, the target and background features are separated by performing contrastive regularization, so that the main branch fully extracts the overall feature of images while converging to the key target features. In addition, the texture and dense maps obtained by processing the input images are used to learn the texture and morphological features contained in the scenes, thus guiding the main branching features to better focus on the unique properties within classes and reduce the impact of sample differences.

A. Target-Background Separation Strategy

Remote sensing scene labels are usually determined by the internal key targets, so most previous methods try to extract the key target information from the input image. However, actual scenes also contain some regions and targets unrelated to the scene label, which are called background here. As shown in Fig. 1 (a), due to the target similarity between scenes, the key target features between different classes are prone to confusion, while the difference of background information provides decision assistance for sample distinction, thus enhancing the discriminative of inter-class samples with similar targets. From the samples as shown in Fig. 1 (b), it can be seen that the background information alone also has inter-class similarity problems, and at this time the difference of key targets provides sufficient information for class judgment. Previous related works try to separate the target from the background and use contextual information, however, their directions are mostly focused on segmentation and detection [34] [35] [36], and there is no complete related work in the field of remote sensing scene classification before. To this end, a target-background separation strategy is proposed to extract the background and key target information contained in the input image respectively. The whole process is illustrated in Fig. 3.

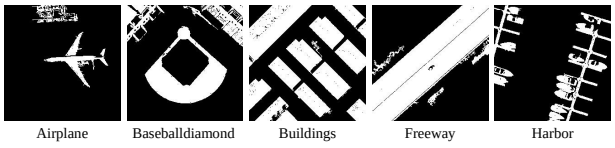


Fig. 4. Segmentation results of images from different classes.

In order to obtain the effective target and background information in the scene respectively, a corresponding mask map is designed first, which can be calculated as:

$$M = Adapthres(X_{gray}) \tag{1}$$

where $Adapthres$ is the adaptive threshold segmentation function and X_{gray} is the grayscale map converted from the original image. To better demonstrate the segmentation results, several images from different classes are selected and their segmented results are shown in Fig. 4.

After that, the original image X and the blurred image $X_{blurred}$ obtained by Gaussian blurring are filled in the corresponding area according to the segmentation result, so as to avoid the influence of other information on the feature extraction of specific branches and make the background and target branches focus on the corresponding features in the image respectively,

$$\begin{aligned} X_{target} &= \begin{cases} X & M \neq 0 \\ X_{blurred} & M == 0 \end{cases} \\ X_{background} &= \begin{cases} X_{blurred} & M \neq 0 \\ X & M == 0 \end{cases} \end{aligned} \tag{2}$$

where X_{target} and $X_{background}$ are the inputs of the target and background branches.

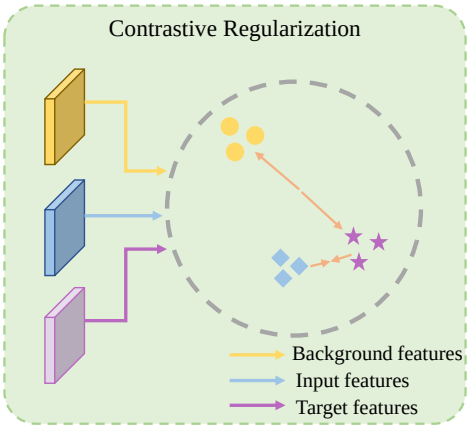


Fig. 5. The contrastive regularization is adopted to pull the main branch features to the target features and push the background features away the target features.

B. Contrastive Regularization

Inspired by contrastive learning, it aims to learn a feature representation in a certain feature space that draws the "positive" pairs closer and pulls the "negative" pairs apart. To this end, an contrastive regularization (CR) is introduced to normalize and generate better feature representations, which is shown in Fig. 5. There are three feature extraction branches in MGSNet, the background and target branches are used to extract the separated background and target features, and the main branch is used to mine the overall features of the scene image. To this end, the positive and negative pairs in CR are features extracted from the target F_t and main branches F_m and features from the background F_b and target branches, respectively. Therefore, the loss function of the network can be reformulated as:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} -y_i^c \log(p_i^c) + \beta\rho(F_b, F_t, F_m) \tag{3}$$

where the first term is the cross-entropy loss between the true label and the output of MGSNet to regularize the learning of the classifier. N is the number of input samples, C is the number of scene classes, and y is the corresponding

label; if the class of input is c, then y^c=1, otherwise 0; p^c is the prediction probability after softmax. The second term $\rho(F_b, F_t, F_m)$ represents the contrastive regularization of F_b, F_t and F_m under the same latent feature space, which plays an important role in bringing the main branch features closer to the target features and the background features far away the target features. β is a hyperparameter used to balance the cross-entropy loss and CR. To enhance the contrastive power between features, hidden features from multistage are extracted and subjected to contrastive regularization. Thus, the overall loss function can be further calculated as:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} -y_i^c \log(p_i^c) + \beta \sum_{j=1}^{h} w_j \frac{D(F_m^j, F_t^j)}{D(F_b^j, F_t^j)} \quad (4)$$

where h is the number of hidden layers, and $j = 1, 2, \cdots h$ is the jth hidden layer from the model. $D(x, y)$ is the L1 loss between x and y, w_j is the weight coefficient.

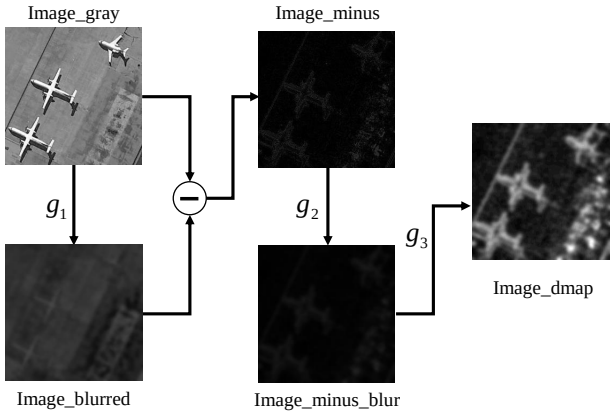


Fig. 6. Illustration of computing the dense map.

C. Self-Guided Network

As mentioned before, due to the varied lighting conditions, view angle, and the complexity of ground surfaces, there are still differences in presentation forms between scenes of the same class. To this end, the designed self-guided network reduces the impact of performance differences by exploiting the inherent structural features of samples to guide feature extraction and learning. It consists of two parts, dense map guidance and texture structure guidance, starting from morphological features and texture features respectively, to improve the intra-class consistency of sample features.

1) Dense Map Guidance: In general, scene images contain complex regions with dense semantic information and smooth regions with low information frequency, and the mixing of these regions poses a challenge for the model to focus on the key semantic regions in the image. Therefore, it is suboptimal to blindly process the entire input sample in the same way. Inspired by the success of noise map in denoising task, a dense map X_D is introduced here to guide the network to perform targeted processing on different regions of the scene. In the dense map, regions with dense semantic information correspond to high-frequency patterns, while regions with less texture correspond to low-frequency patterns. It can be calculated as:

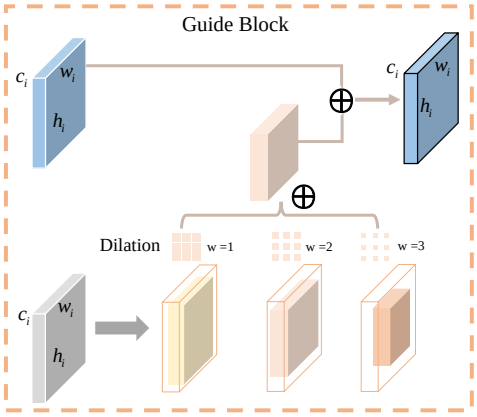


Fig. 7. Guide block with different dilation rates for texture structure guidance.

$$X_D = g_3(g_2(I_{gray} - g_1(I_{gray}; K_1) : K_2)) \quad (5)$$

where both g_1 and g_2 are Gaussian blur operations with the kernel size of K_1 and K_2, g_3 is a normalization function, and I_{gray} is the mean of three channels of the input image. g_3 is defined as:

$$g_3(I) = \frac{I - min(I)}{max(I) - min(I) + \varepsilon} \quad (6)$$

After applying Gaussian blur operation g_1 to the input image, the regions with dense semantic information become blurred, which makes the gap between the output and the original input more obvious. After that, another Gaussian blurring operation g_2 is applied to obtain a smooth density map. Finally, the image is regularized to the interval from 0 to 1. The whole process is shown in Fig. 6.

Fig. 8. Sample images of the UCM dataset: two images of each class are exhibited. (Semantic category ∼ Number of samples.)

Once the dense map is computed, similar to [37], it is combined by concatenating it with other channels, and then fed into the network as the input to the main branch.

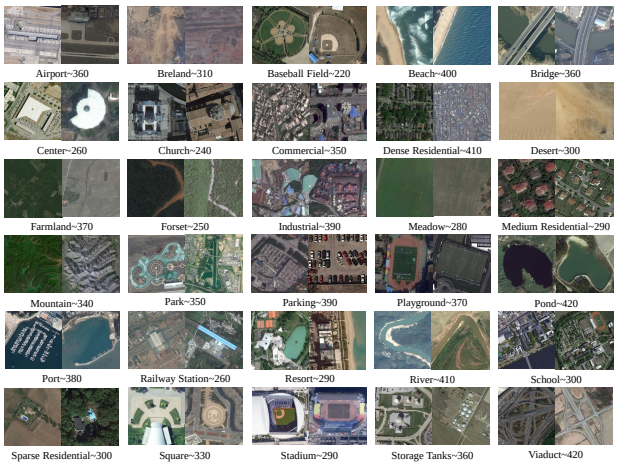


Fig. 9. Sample images of the AID dataset: two images of each class are exhibited. (Semantic category $\sim$ Number of samples.)

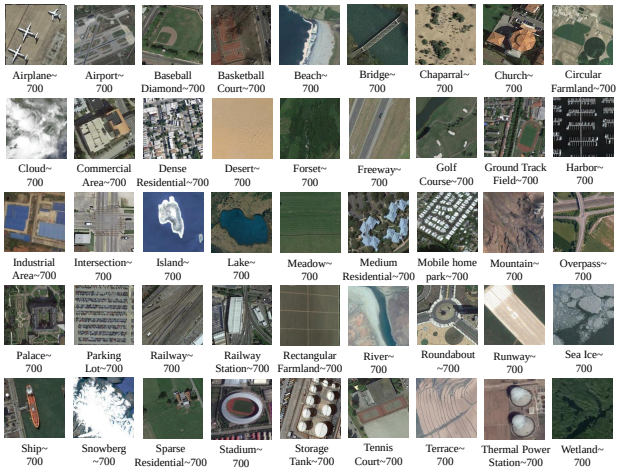


Fig. 10. Sample images of the NWPU-RESISC45 dataset: one image of each class are exhibited. (Semantic category $\sim$ Number of samples.)

2) Texture Structure Guidance: Due to the different acquisition times and methods of intra-class samples, their representation forms are diverse, while the inherent texture structure features in the data provide a good guide for feature extraction. For this reason, a texture structure guidance module is proposed, and its structure is shown in Fig. 7.

The gray image obtained from the original input discards the influence caused by color and light intensity and focuses on the texture structure information. The abstract features extracted by different stages are sent to the texture structure guidance module together with the main branch features. Within the module, the features with different receptive fields are obtained through convolution operations with different dilation rates, while effectively modeling the scale variation associated with the object texture. Specifically, let F_m and F_d denote the features extracted by the main branch and the texture features at the same level, respectively. By defining different dilation rates $d = 1, 2, 3$, F_d can obtain features with different receptive fields. Combining these features with main branch features can better model scene features of different

TABLE I
COMPARISON OF OVERALL ACCURACY AND STANDARD DEVIATIONS (%) OF STATE-OF-THE-ART METHODS ON UCM DATASET WITH THE TRAINING RATIO OF 80%

Type	Method	Publication Year	Training ratio 80% (20% testing)
†	BoVW(LBP) [38]	TGRS2017	77.12±1.93
	BoVW(SIFT) [38]	TGRS2017	74.12±3.30
	salM^{3}LBP-CLM [39]	JSTARS2017	95.75±0.80
	salCLM(eSIFT) [39]	JSTARS2017	94.52±0.79
‡	Two-Fusion [40]	CIN2018	98.02±1.03
	CCPNet [41]	RS2018	97.52±0.97
	GCFs+LOFs [42]	RS2018	99.00±0.35
	CNN-CapsNet [43]	RS2019	99.05±0.24
	SCCov [44]	TNNLS2019	99.05±0.25
	ARCNet-VGG [45]	TGRS2019	99.12±0.40
	GBNet [46]	TGRS2020	98.57±0.48
	MG-CAP [47]	TIP2020	99.00±0.10
	BiMobileNet [48]	Sensors2020	99.03±0.28
	SEMSDNet [49]	JSTARS2021	99.41±0.41
	CSDS [50]	JSTARS2021	99.52±0.13
	T-CNN [51]	TGRS2022	99.33±0.11
	DFAGCN [13]	TNNLS2022	98.48±0.42
Ours	MGSNet		OA: **99.76±0.14** AA: 99.76±0.14 KC: 99.75±0.16

†:Handcrafted Feature-Based Methods ‡:Deep Feature-Based Methods

scales. This process can be expressed as:

$$F_m = F_m \oplus \sum_d Conv(F_d, d) \tag{7}$$

With the guidance of multiscale texture features, the feature extraction and learning of the main branch better deal with the inter-class samples with different representations and improve the unity of feature representation.

D. Analysis on the Proposed Method

This paper proposes a target-background separation strategy and contrastive regularization for the insufficient feature representation caused by the difference of target-background imbalance changes in scene classification tasks. In addition, in the face of the problem of intra-class sample variability, the self-guided network is used to guide the feature learning of the network branch.

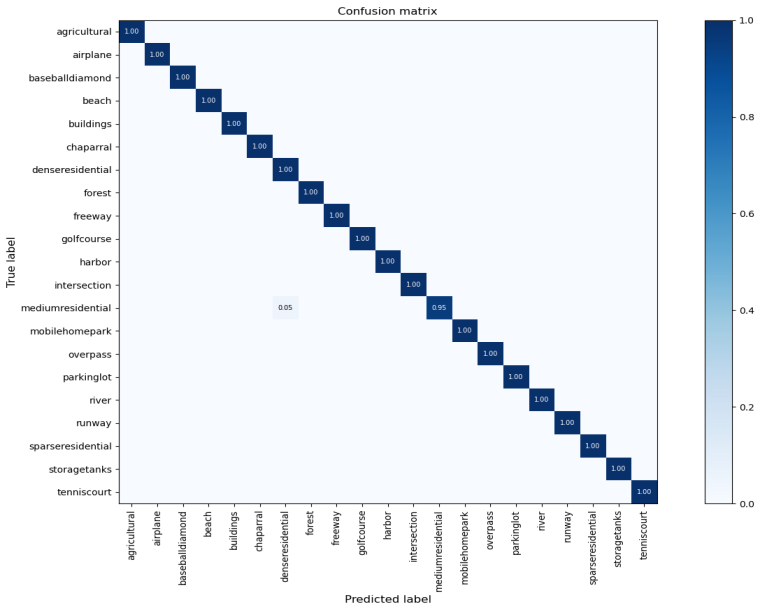
The motivation of this method is to use the background information ignored by existing methods as an aid to scene classification, so as to improve the discrimination of the model for input samples. To achieve this goal and reduce the interference of irrelevant information, a unique separation strategy is proposed and verified. Only a suboptimal separation can be achieved only by processing the input samples alone. Therefore, a contrastive regularization method is introduced to reduce the interference of irrelevant information under the specific branch, and its enhancement for model understanding is discussed in Section III-D.

Furthermore, to improve the grasp of the overall information of the scene and eliminate the interference generated during the sample collection process, a self-guided network came into being. Its goal is to utilize the uniform texture and

TABLE II
COMPARISON OF OVERALL ACCURACY AND STANDARD DEVIATIONS (%) OF STATE-OF-THE-ART METHODS ON AID DATASET WITH THE TRAINING RATIO OF 20% AND 50%

Type	Method	Publication Year	Training ratios	
			20%(80% testing)	50%(50% testing)
†	BoVW(LBP)) [38]	TGRS2017	56.73±0.54	64.25±0.55
	BoVW(SIFT) [38]	TGRS2017	61.40±0.41	67.65±0.49
	salM^{3}LBP-CLM [39]	JSTARS2017	86.92±0.35	89.76±0.45
	salCLM(eSIFT) [39]	JSTARS2017	85.58±0.83	88.41±0.63
‡	Two-Fusion [40]	CIN2018	92.32±0.41	94.58±0.25
	GCFs+LOFs [42]	RS2018	92.48±0.38	96.85±0.23
	CNN-CapsNet [43]	RS2019	93.79±0.13	96.32±0.12
	SCCov [44]	TNNLS2019	93.12±0.25	96.10±0.16
	ARCNet-VGG [45]	TGRS2019	88.75±0.40	93.10±0.55
	GBNet [46]	TGRS2020	92.20±0.23	95.48±0.12
	MG-CAP [47]	TIP2020	93.34±0.18	96.12±0.12
	BiMobileNet [48]	Sensors2020	94.83±0.24	96.87±0.23
	CSDS [50]	JSTARS2021	94.29±0.35	96.70±0.14
	PSGAN [52]	TGRS2022	89.47±0.34	92.67±0.55
	T-CNN [51]	TGRS2022	94.55±0.27	96.27±0.23
Ours	MGSNet		OA: **95.46±0.21** AA: 95.25±0.27 KC: 95.60±0.31	OA: **97.18±0.16** AA: 97.12±0.19 KC: 97.08±0.22

†:Handcrafted Feature-Based Methods ‡:Deep Feature-Based Methods



CM of UCM dataset (Training ratio = 80%)

Fig. 11. Confusion matrix (CM) on UCM dataset with 80% of the dataset for training and the rest for testing.

topological features of the intra-class samples to compensate for the differences in representation forms, and guide the learning of main branch features. Specifically, the combination of the dense map gives extra attention to the key area of the image, and the texture structure information at different scales is captured through the convolution operation with different dilation rates. The combination of them effectively promotes the mining of the essential features of the image.

III. EXPERIMENTAL RESULTS AND DISCUSSIONS

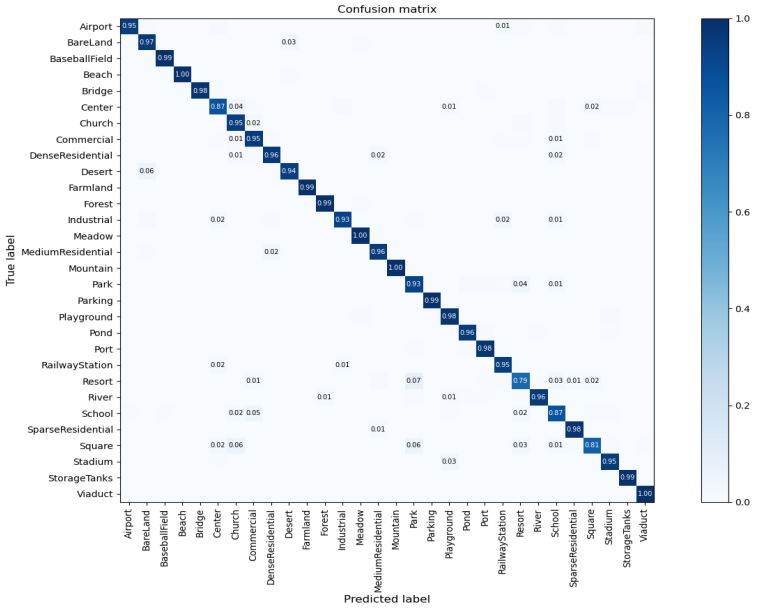
A. Datasets Descriptions and Evaluation Metrics

1) Datasets Descriptions: (1) A series of experiments are performed on the UCM dataset [54], which is one of the most classic datasets in remote sensing image scene classification tasks. It consists of 21 scene classes, including agricultural, airplane, baseballdiamond, beach, etc. It is derived from the United States Geological Survey National Map, Urban Area Imagery collection, where each class consists of 100 images with a size of 256× 256 pixels, which together form a scene

TABLE III
COMPARISON OF OVERALL ACCURACY AND STANDARD DEVIATIONS (%) OF STATE-OF-THE-ART METHODS ON NWPU-RESISC45 DATASET WITH THE TRAINING RATIO OF 10% AND 20%

Type	Method	Publication Year	Training ratios	
			10%(90% testing)	20%(80% testing)
†	BoVW [12]	RPOC2017	41.72±0.21	44.79±0.28
	BoVW+SPM [12]	RPOC2017	27.83±0.61	32.96±0.47
	LLC [12]	RPOC2017	38.81±0.23	40.03±0.34
‡	Fine-tuned VGG-16 [12]	RPOC2017	87.15±0.45	90.36±0.18
	Two-Fusion [40]	CIN2018	80.22±0.22	83.16±0.18
	CNN-CapsNet [43]	RS2019	89.03±0.21	92.60±0.11
	SCCov [44]	TNNLS2019	89.30±0.35	92.10±0.25
	MF^2Net [53]	GRSL2020	90.17±0.25	92.73±0.21
	MG-CAP [47]	TIP2020	90.83±0.12	92.95±0.11
	BiMobileNet [48]	Sensors2020	92.06±0.14	94.08±0.11
	SEMSDNet [49]	JSTARS2021	91.68±0.39	93.89±0.63
	CSDS [50]	JSTARS2021	91.64±0.16	93.59±0.21
	PSGAN [52]	TGRS2022	84.72±0.72	88.47±0.56
	T-CNN [51]	TGRS2022	90.25±0.14	93.05±0.12
Ours	MGSNet		OA: **92.40±0.16** AA: 92.40±0.16 KC: 92.23±0.18	OA: **94.57±0.12** AA: 94.57±0.12 KC: 94.44±0.15

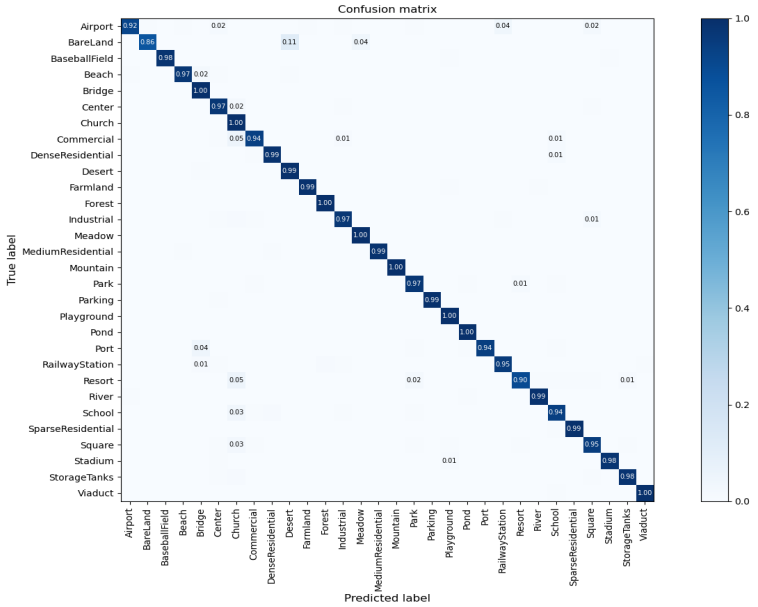
†:Handcrafted Feature-Based Methods ‡:Deep Feature-Based Methods



CM of AID dataset (Training ratio = 20%)

Fig. 12. Confusion matrix (CM) on AID dataset with 20% of the dataset for training and the rest for testing.

classification dataset containing 2100 images. Fig. 8 shows some samples of this benchmark dataset. To achieve a fair comparison, 80% of the images are randomly selected as the training set, and the remaining images are used for testing. (2) The AID dataset is used in the experiment [38], which is a large-scale dataset consisting of 10,000 aerial scenes images with a size of 600 × 600 pixels, containing 30 scene categories such as airport, bare land, baseball field, beach, bridge, etc. All images were collected via Google Earth and annotated by experts, with spatial resolution changes from half a meter to 8m. Fig. 9 illustrates some images of each class in this dataset. Similar to UCM, 20% and 50% of the samples are randomly selected as the training set to optimize the model parameters. (3) The NWPU-RESISC45 dataset is a large-scale dataset created by Northwestern Polytechnic University using Google Earth Imagery [12], which contains 45 scene categories including airplane, airport, baseball diamond, basketball court, beach, etc. Each class consists of 700 images with a size of 256 × 256, and its spatial resolution ranging from 0.2 m to 30 m. Fig. 10 exhibits some images of this challenging dataset. 10% and 20% of the images are selected as the training set, and the remaining images are used to test the performance of



CM of AID dataset (Training ratio = 50%)

Fig. 13. Confusion matrix (CM) on AID dataset with 50% of the dataset for training and the rest for testing.

the model.

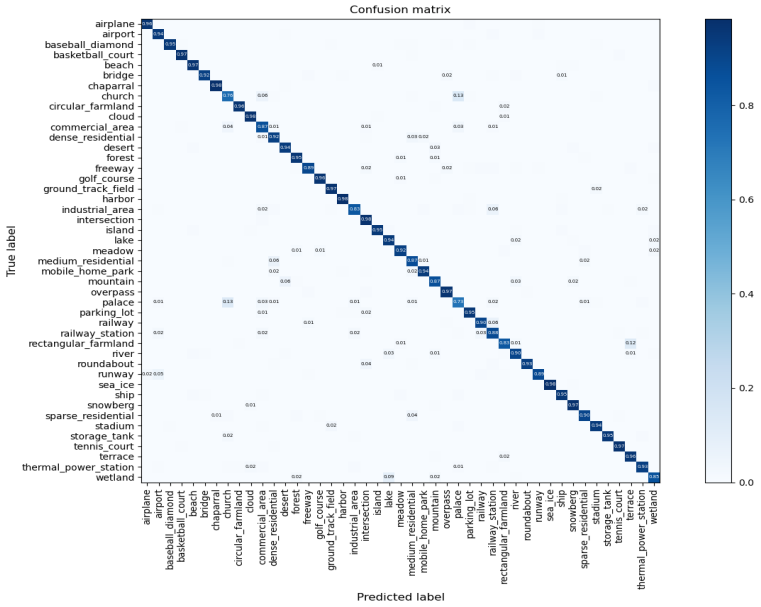
2) Evaluation Metrics: For better quantitative analysis of the experimental results, two widely used evaluation metrics were introduced. (1) Overall accuracy (OA). The overall accuracy is the ratio between the number of correctly classified samples and the overall samples. (2) Average accuracy (AA). The average accuracy refers to the average of each class accuracy, which reflects the balance of the classification results. (3) Kappa coefficient (KC). kappa coefficient an indicator for consistency detection, which is used to measure the effectiveness of the classification. In addition, the confusion matrix is introduced to visualize the inter-class classification error and the degree of confusion, thus enabling qualitative analysis of the classification results. Specifically, each column represents the prediction result of the model, the sum of each column represents the number of samples predicted to be that class, each row represents the true distribution, and the sum of the data in each row represents the true number of samples in that class. In this way, it shows the correct classification and the misclassification of the model, which can help readers better understand the performance of the model.

B. Comparison with State-of-the-art Approaches

In this subsection, the experimental results of the three datasets are described and discussed in detail. To verify the effectiveness of the proposed method, some state-of-the-art methods from recent years are used for comparison. The comparison results on UCM, AID, and NWPU-RESISC45 data sets are reported in Table I, Table II and Table III, respectively.

Table I lists the classification results of the proposed method and other state-of-the-art methods on the UCM dataset. In order to prevent the undue influence of the experimental samples on the experimental results and to maintain the fairness of the results of different methods, all methods take 80% of the samples as the training set for model optimization. Judging from the classification accuracy of each method in the table, the understanding ability of handcrafted feature-based methods limits the classification accuracy of the model due to the limitation of feature extraction, thus affecting the final classification results. While the classification accuracy of most deep feature-based methods can reach 90% and then tend to be stable and difficult to improve. MGSNet achieves a higher classification accuracy of 99%, which is a significant improvement compared to other methods. This further demonstrates that the proposed MGSNet not only effectively alleviates the difference caused by the target-background imbalance change, but also improves the feature consistency of the intra-class samples, which is conducive to the improvement of classification accuracy.

The classification performance of the comparison methods and MGSNet on the AID dataset are listed in Table II, where two columns of results represent using 20% and 50% samples for training, and the remaining samples as the test set. Compared with handcrafted feature-based methods, deep feature-based methods have an improvement of at least 1.83% and 3.34% when the proportion of training samples is 20% and 50%, with a great improvement. The improvement also fully demonstrates the benefit of deep feature extraction. In addition, with the increase of the number of training samples, the classification accuracy has a greater gap, which indicates the demand for training samples of deep feature-based methods. The proposed MGSNet achieves the best classification performance of 95.46% and 97.18% respectively under different training ratios and has improved the accuracy of 8.54%, 7.42% and 0.63%, 0.31% compared to the two types of methods, which confirms the effectiveness of the proposed method on



CM of NWPU-RESISC45 dataset (Training ratio = 10%)

Fig. 14. Confusion matrix (CM) on NWPU-RESISC45 dataset with 10% of the dataset for training and the rest for testing.

TABLE IV
ABLATION STUDIES FOR THE PROPOSED MGSNET ON THREE DATASETS

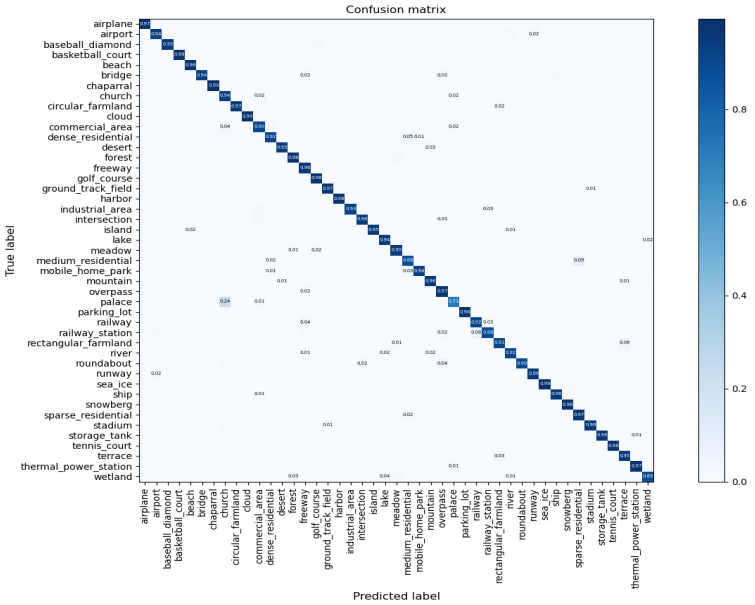
Variant	Target-background separation strategy		CR	Self-guided network		OA (%) on different dataset				
	Target branch	Background branch		Dense map guidance	Texture structure guidance	UCM(80%)	AID(20%)	AID(50%)	NWPU(10%)	NWPU(20%)
1	✓	✓	✓	✓	✓	99.76	95.46	97.18	92.40	94.57
2	✓	✓		✓	✓	99.52	95.20	96.92	92.11	94.28
3		✓		✓	✓	99.36	94.98	96.60	91.87	93.99
4	✓			✓	✓	99.28	94.73	96.41	91.66	93.72
5				✓	✓	99.04	94.31	94.18	91.32	93.40
6	✓	✓	✓	✓		99.33	94.72	96.45	91.82	93.91
7	✓	✓	✓		✓	99.56	95.02	96.84	92.04	94.15
8	✓	✓	✓			99.20	94.47	96.21	91.66	93.64
9	✓	✓				99.01	94.23	95.98	91.41	93.39

the AID dataset.

Compared with the first two datasets, the NWPU-RESISC45 dataset has the largest number of scene classes as well as the amount of data, so its rich image variations, differences of intra-class samples, and similarities between inter-class samples make the dataset more challenging. The comparison results between the proposed method and existing methods using the NWPU-RESISC45 dataset are listed in Table III. With 10% and 20% of the samples are respectively chosen as the training set, and the remaining samples are used for testing, the proposed method outperforms other comparative methods. When the training rate is 10%, the overall accuracy of MGSNet is 2.15% higher than the latest research T-CNN based on depth feature, and 50.68% higher than the methods based on handcrafted feature. With a 20% training share, there is an improvement of 1.52% and 49.78%. It can be seen that the learning ability and discrimination of the model are effectively improved by introducing background information and improving the consistency of intra-class samples.

C. Reports of Confusion Matrix

To better demonstrate the specific classification performance of the proposed method, confusion matrices are drawn to illustrate the classification results. A total of five confusion matrices can be obtained through experimental results with different training ratios on the UCM, AID, and NWPU-RESISC45 datasets, which are presented in Fig.11-Fig.15. Each column of the confusion matrix represents the predicted label, and the sum of each column represents the number of samples predicted to be the corresponding scene. Each row represents the actual label of the sample, and the sum of each row represents the number of samples of the corresponding scene in the test set. Thus, the cells on the diagonal represent the proportion of samples that are correctly classified, while the other cells are misclassified cases.



CM of NWPU-RESISC45 dataset (Training ratio = 20%)

Fig. 15. Confusion matrix (CM) on NWPU-RESISC45 dataset with 20% of the dataset for training and the rest for testing.

Fig.11 shows the confusion matrix when the training set accounted for 80% on the UCM dataset. Among the 21 scene classes, only one scene has an accuracy below 100%, reaching 95%, which is the misclassification of *medium residential* into *dense residential*. This is because the target and background information in these two types of scenes are too similar, so it is easy to cause model misclassification when processing relevant samples. However, MGSNet still has unique advantages, for example, when dealing with *freeway*, *runway*, and *overpass*, although the valid targets are all roads, the differences of their background information can be effectively captured by the network to correctly distinguish different classes.

Fig.12 and Fig.13 show the confusion matrices obtained on the AID dataset when the training set accounts 20% and 50%. Only four and one scene classes' accuracy is less than 90%, and the rest classes have achieved excellent performance. For example, the *center* has more misclassifications because its object shape and context information are similar to the *church*. However, the problem of intra-class variation can be better resolved in the division of classes such as *forest* due to the introduction of texture and morphological information for the guidance of network learning.

The confusion matrices of NWPU-RESISC45 dataset are shown in Fig.14 and Fig.15. When the proportion of training samples is 10% and 20%, there are 34 and 41 scenes with classification accuracy exceeding 90%. Among the two groups of experiments, the worst classification performance is *palace*, which is wrongly classified as *church* in most cases. This is because there are certain similarities in valid targets, background, and context information, which makes the model prone to misclassification. On the contrary, although *desert* has a certain similarity with *lake* in the target performance, MGSNet can recognize it well with the assistance of background information. Similarly, the similarity between *chaparral* and *desert* backgrounds can be effectively distinguished by supplementing the information of valid targets, thereby improving the robustness of the model.

D. Ablation Studies

To verify the role of separation strategy, contrastive regularization, and self-guided network in this task, this subsection conducts a series of ablation experiments to explore the importance of target-background information for network performance improvement and the effectiveness of various parts of self-guided network. The specific experimental results are listed in Table IV. Specifically, the whole ablation experiments can be divided into two parts: 1) Verify the effect of target-background information and contrastive regularization on improving feature discrimination. Through the comparison of #1 and #2, the introduction of contrastive regularization increases the separation between different features and improves the extraction ability of specific features by each network branch. And the comparisons between#2, #3, #4 and #5 fully demonstrate that the additional introduction of target-background information is helpful in improving the discrimination between similar scenes for better inter-class sample differentiation. While #9 fully demonstrates the experimental performance only using the variant with target branch and background branch only. It can be seen that even without adding other modules, good results can still be achieved through the separation and support of the target and background information. 2) Verify the assistance of self-guided network for main branch feature extraction. Through the comparison of the results of #1, #6, #7, and #8, it can clearly see the benefits of dense graph guidance and texture structure guidance for improving the consistency of the main branch feature representation. They utilize the morphology and texture features of images to

guide the feature attention, reducing the impact of differences in intra-class sample representations (e.g., seasonal turnover, regional variation), and allowing the model to focus more on common features within the samples. By combining the target-background separation strategy, contrastive regularization and self-guided network, the proposed MGSNet effectively alleviates the misclassification phenomenon caused by the inter-class similarity and the intra-class differences, making it achieve state-of-the-art performance in multiple datasets.

E. Running time and Memory Requirement

To demonstrate the running efficiency and memory requirement of the model, relevant statistics are performed on the UCM dataset. The training time of the model is calculated with 32 images within a round of training, while the test time is obtained using the entire test set (420 images). The training time is 0.307s and the testing time is 0.614s, which is the average of the five runs of the model. In addition, the memory requirement of the model is 7.58 MB. From the obtained results, the model runs efficiently and requires a moderate amount of memory, which is sufficient for subsequent applications.

IV. Conclusions

In this paper, a novel MGSNet has been proposed for remote sensing scene classification. First of all, to solve the problem of inter-class sample similarity caused by the unbalanced change of target-background, a target-background separation strategy, and a contrastive regularization method are established in MGSNet. Specifically, the input samples are divided into target and background regions by separating them. Afterwards, they are fed into the corresponding feature extraction branches to achieve separate extraction of target and background information, thus providing more and more reliable information assistance for sample differentiation. And contrastive regularization serves this strategy, improving the feature separation by expanding the difference between them, so that each branch pays more attention to the extraction and utilization of its corresponding information. In addition, a self-guided network based on image texture and morphological features is proposed to cope with differentiated intra-class samples. With the guidance of dense map and texture structure information, the main branch pays more attention to the common characteristics of samples in the class. Extensive experimental results on three datasets demonstrate the superiority of MGSNet from the perspective of quantitative indicators and confusion matrix, proving its effectiveness on remote sensing scene classification tasks.

References

[1] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 6, pp. 3660–3671, 2016.

[2] Y. Gao, M. Zhang, W. Li, X. Song, X. Jiang, and Y. Ma, "Adversarial complementary learning for multisource remote sensing classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–13, 2023.

[3] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1560–1584, 2015.

[4] J. Wang, W. Li, Y. Gao, M. Zhang, R. Tao, and Q. Du, "Hyperspectral and sar image classification via multiscale interactive fusion network," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2022.

[5] X. Huang, H. Liu, and L. Zhang, "Spatiotemporal detection and analysis of urban villages in mega city regions of china using high-resolution remotely sensed imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 7, pp. 3639–3657, 2015.

[6] Z. Yu, Z. Liu, Y. Zhang, Y. Qu, and C.-Y. Su, "Distributed finite-time fault-tolerant containment control for multiple unmanned aerial vehicles," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 6, pp. 2077–2091, 2020.

[7] W. Li, J. Wang, Y. Gao, M. Zhang, R. Tao, and B. Zhang, "Graph-feature-enhanced selective assignment network for hyperspectral and multispectral data classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[8] T.-Z. Xiang, G.-S. Xia, and L. Zhang, "Mini-unmanned aerial vehicle-based remote sensing: Techniques, applications, and prospects," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 3, pp. 29–63, 2019.

[9] W. Li, Y. Gao, M. Zhang, R. Tao, and Q. Du, "Asymmetric feature fusion network for hyperspectral and sar image classification," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–14, 2022.

[10] J. Kang, R. Fernandez-Beltran, Z. Ye, X. Tong, P. Ghamisi, and A. Plaza, "Deep metric learning based on scalable neighborhood components for remote sensing scene characterization," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 12, pp. 8905–8918, 2020.

[11] 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 3735–3756, 2020.

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

[13] 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 Transactions on Neural Networks and Learning Systems*, vol. 33, no. 10, pp. 5751–5765, 2022.

[14] H. Li, H. Gu, Y. Han, and J. Yang, "Object-oriented classification of high-resolution remote sensing imagery based on an improved colour structure code and a support vector machine," *International journal of remote sensing*, vol. 31, no. 6, pp. 1453–1470, 2010.

[15] E. Aptoula, "Remote sensing image retrieval with global morphological texture descriptors," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 5, pp. 3023–3034, 2014.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, pp. 91–110, 2004.

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 886–893 vol. 1.

[18] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 4, pp. 2175–2184, 2015.

[19] X. Lu, X. Zheng, and Y. Yuan, "Remote sensing scene classification by unsupervised representation learning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 9, pp. 5148–5157, 2017.

[20] J. Wang, F. Gao, J. Dong, S. Zhang, and Q. Du, "Change detection from synthetic aperture radar images via graph-based knowledge supplement network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 1823–1836, 2022.

[21] Y. Gao, W. Li, M. Zhang, J. Wang, W. Sun, R. Tao, and Q. Du, "Hyperspectral and multispectral classification for coastal wetland using depthwise feature interaction network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[22] M. Zhao, W. Li, L. Li, J. Hu, P. Ma, and R. Tao, "Single-frame infrared small-target detection: A survey," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 2, pp. 87–119, 2022.

[23] Y. Gao, M. Zhang, J. Wang, and W. Li, "Cross-scale mixing attention for multisource remote sensing data fusion and classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2023.

[24] H. He and H. Jiang, "Deep learning based energy efficiency optimization for distributed cooperative spectrum sensing," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 32–39, 2019.

[25] J. Fang, Y. Yuan, X. Lu, and Y. Feng, "Robust space–frequency joint representation for remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 10, pp. 7492–7502, 2019.

[26] M. Zhang, W. Li, Y. Zhang, R. Tao, and Q. Du, "Hyperspectral and lidar data classification based on structural optimization transmission," *IEEE Transactions on Cybernetics*, pp. 1–12, 2022.

[27] 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 Transactions on Image Processing*, vol. 29, pp. 4911–4926, 2020.

[28] Z. Zhao, J. Li, Z. Luo, J. Li, and C. Chen, "Remote sensing image scene classification based on an enhanced attention module," *IEEE Geoscience and Remote Sensing Letters*, vol. 18, no. 11, pp. 1926–1930, 2021.

[29] Y. Liu, Y. Zhong, and Q. Qin, "Scene classification based on multiscale convolutional neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 12, pp. 7109–7121, 2018.

[30] 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 Transactions on Geoscience and Remote Sensing*, vol. 56, no. 5, pp. 2811–2821, 2018.

[31] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, "Integrating multilayer features of convolutional neural networks for remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5653–5665, 2017.

[32] S.-B. Chen, Q.-S. Wei, W.-Z. Wang, J. Tang, B. Luo, and Z.-Y. Wang, "Remote sensing scene classification via multi-branch local attention network," *IEEE Transactions on Image Processing*, vol. 31, pp. 99–109, 2022.

[33] H. Wu, Y. Qu, S. Lin, J. Zhou, R. Qiao, Z. Zhang, Y. Xie, and L. Ma, "Contrastive learning for compact single image dehazing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 551–10 560.

[34] M. D. Hossain and D. Chen, "Segmentation for object-based image analysis (obia): A review of algorithms and challenges from remote sensing perspective," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 150, pp. 115–134, 2019.

[35] X. Feng, J. Han, X. Yao, and G. Cheng, "Progressive contextual instance refinement for weakly supervised object detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 8002–8012, 2020.

[36] W. Zhou, J. Liu, J. Lei, L. Yu, and J.-N. Hwang, "Gmnet: Graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 7790–7802, 2021.

[37] M. Gharbi, G. Chaurasia, S. Paris, and F. Durand, "Deep joint demosaicking and denoising," *ACM Transactions on Graphics (ToG)*, vol. 35, no. 6, pp. 1–12, 2016.

[38] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "Aid: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965–3981, 2017.

[39] X. Bian, C. Chen, L. Tian, and Q. Du, "Fusing local and global features for high-resolution scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 6, pp. 2889–2901, 2017.

[40] Y. Yu and F. Liu, "A two-stream deep fusion framework for high-resolution aerial scene classification," *Computational intelligence and neuroscience*, vol. 2018, 2018.

[41] K. Qi, Q. Guan, C. Yang, F. Peng, S. Shen, and H. Wu, "Concentric circle pooling in deep convolutional networks for remote sensing scene classification," *Remote Sensing*, vol. 10, no. 6, p. 934, 2018.

[42] D. Zeng, S. Chen, B. Chen, and S. Li, "Improving remote sensing scene classification by integrating global-context and local-object features," *Remote Sensing*, vol. 10, no. 5, p. 734, 2018.

[43] W. Zhang, P. Tang, and L. Zhao, "Remote sensing image scene classification using cnn-capsnet," *Remote Sensing*, vol. 11, no. 5, p. 494, 2019.

[44] N. He, L. Fang, S. Li, J. Plaza, and A. Plaza, "Skip-connected covariance network for remote sensing scene classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 5, pp. 1461–1474, 2020.

[45] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of vhr remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 2, pp. 1155–1167, 2019.

[46] H. Sun, S. Li, X. Zheng, and X. Lu, "Remote sensing scene classification by gated bidirectional network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 1, pp. 82–96, 2020.

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

[48] D. Yu, Q. Xu, H. Guo, C. Zhao, Y. Lin, and D. Li, "An efficient and lightweight convolutional neural network for remote sensing image scene classification," *Sensors*, vol. 20, no. 7, p. 1999, 2020.

[49] T. Tian, L. Li, W. Chen, and H. Zhou, "Semsdnet: A multiscale dense network with attention for remote sensing scene classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 5501–5514, 2021.

[50] X. Wang, L. Yuan, H. Xu, and X. Wen, "Csds: End-to-end aerial scenes classification with depthwise separable convolution and an attention mechanism," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 10 484–10 499, 2021.

[51] W. Wang, Y. Chen, and P. Ghamisi, "Transferring cnn with adaptive learning for remote sensing scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–18, 2022.

[52] G. Cheng, X. Sun, K. Li, L. Guo, and J. Han, "Perturbation-seeking generative adversarial networks: A defense framework for remote sensing image scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[53] K. Xu, H. Huang, Y. Li, and G. Shi, "Multilayer feature fusion network for scene classification in remote sensing," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 11, pp. 1894–1898, 2020.

[54] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, 2010, pp. 270–279.