GCFnet: Global Collaborative Fusion Network for Multispectral and Panchromatic Image Classification

Hui Zhao, Sicong Liu, *Senior Member, IEEE*, Qian Du, *Fellow, IEEE*, Lorenzo Bruzzone, *Fellow, IEEE*, Yongjie Zheng, Kecheng Du, Xiaohua Tong, *Senior Member, IEEE*, Huan Xie, *Senior Member, IEEE*, Xiaolong Ma

***Abstract*—**Among various multimodal remote sensing data, the pairing of multispectral (MS) and panchromatic (PAN) images is widely used in remote sensing applications. This article proposes a novel global collaborative fusion network (GCFnet) for joint classification of MS and PAN images. In particular, a global patch-free classification scheme based on an encoder-decoder deep learning (DL) network is developed to exploit context dependencies in the image. The proposed GCFnet is designed based on a novel collaborative fusion architecture, which mainly contains three parts: 1) two shallow-to-deep feature fusion branches related to individual MS and PAN images; 2) a multiscale cross-modal feature fusion branch of the two images, where an adaptive loss weighted fusion strategy is designed to calculate the total loss of two individual and the cross-modal branches; 3) a probability weighted decision fusion strategy for the fusion of the classification results of three branches to further improve the classification performance. Experimental results obtained on three real datasets covering complex urban scenarios confirm the effectiveness of the proposed GCFnet in terms of higher accuracy and robustness compared to existing methods. By utilizing both sampled and non-sampled position data in the feature extraction process, the proposed GCFnet can achieve excellent performance even in a small sample-size case. The codes will be available from the website: https://github.com/SicongLiuRS/GCFnet.

***Index Terms*—Classification, deep learning, global collaborative fusion, feature fusion, remote sensing.**

This work was supported in part by the National Key R&D Program of China under Grant 2018YFB0505000, in part by the National Natural Science Foundation of China under Grants 42071324, 42001387 and by the Shanghai Rising-Star Program (21QA1409100) (*Corresponding author: Sicong Liu*)

Hui Zhao, Sicong Liu, Yongjie Zheng, Kecheng Du, Xiaohua Tong and Huan Xie are with the College of Surveying and Geoinformatics, Tongji University, Shanghai, 200092, China (e-mail: zhaohui@tongji.edu.cn; sicong.liu@tongji.edu.cn; yongjie.Zheng@outlook.com; kecheng_du@163.com; xhtong@tongji.edu.cn; huanxie@tongji.edu.cn).

Qian Du is with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS 39762, USA (e-mail: du@ece.msstate.edu).

Lorenzo Bruzzone is with the Department of Information Engineering and Computer Science, University of Trento, I-38123, Italy (e-mail: lorenzo.bruzzone@unitn.it).

X. Ma is with the Institute of Cartography and Geographic Information System, Chinese Academy of Surveying and Mapping, Beijing 100036, China (e-mail: maxiaolong@casm.ac.cn).

I. Introduction

Nowadays, multimodal remote sensing data, such as multitemporal images [1-3], hyperspectral and LiDAR images [4-6], multispectral (MS) and SAR images [6, 7], MS and panchromatic (PAN) images [8-11], have been widely used for Earth's land-use and land-cover classification and change detection. The fusion of complementary information in the multimodal remote sensing images can enhance the land-object identification accuracy, resulting in better performance than only using a single data source. Among different pairing multimodal remote sensing data, MS and PAN images are the most commonly used and easily accessible data as they are usually simultaneously acquired by many satellite platforms, such as Landsat and SPOT. For some examples of satellites that can acquire high and very-high-resolution (VHR) MS and PAN images, readers can refer to [11]. Despite the low spatial resolution, an MS image usually contains several spectral bands and thus is suitable to separate different classes of land objects. A PAN image contains only one spectral band covering a wide wavelength range, but its high spatial resolution allows an accurate description of the boundaries of objects and their spatial relationships. Therefore, the joint use of MS and PAN images can take full advantage of spectral-spatial information of land objects presented in the two data sources.

In general, there are two categories of fusion methods to integrate MS and PAN images in classification or detection tasks. This first is pixel-level fusion, also known as pan-sharpening (PS) methods, which directly fuses the original MS and PAN images. This has been used in various remote sensing applications, such as change detection and classification [12-14]. However, spectral and spatial distortions may occur in a pan-sharpened image [15]. The other category is feature-level fusion. These methods extract representative features from MS and PAN images and then fuse such features for classification or detection. However, the extraction of useful features often requires domain knowledge [16-18]. In recent years, data-driven deep learning (DL) techniques that can automatically learn abstract and robust deep features from the original data have shown to be very promising for dealing with MS and PAN images fusion. In [19], a superpixel-based multiple local regions combined representation network model was proposed to classify MS image; then a PAN image with detailed spatial information was used to modify the classification results. In [8], a stacked autoencoder was used to describe the spectral information of a MS image, and a

convolutional neural network (CNN) was used to capture spatial features from a PAN image; then spectral and spatial features were concatenated and fed into three fully connected layers to obtain the final classification result. In [20], two CNN modules inspired by the VGG model were designed to extract and combine features from MS and PAN images at their original resolution, and then used in land-cover classification. In [21], the FuseNet was proposed to match the resolution of MS and PAN bands in a VHR image using convolutional layers with down-sampling and up-sampling operations. Then, ReuseNet was built on top of the architecture of FuseNet by incorporating recurrent connections [9]. In [22], a dual-branch attention fusion deep network was designed to extract features for multiresolution image classification, which contains a spatial attention module for PAN image and a channel attention module for MS image. In [23], an adaptive hybrid fusion network (AHF-Net) was designed for MS and PAN image classification, which includes an adaptive weighted intensity-hue-saturation strategy for MS and PAN image fusion, and a correlation-based attention feature fusion module. In [11], deep cross-resolution hidden layer features were extracted from MS and PAN images according to an autoencoder-like deep network, then the selected hidden layer features were used for joint classification of two images.

In general, the above feature-level fusion methods were shown to be more effective in classification than pixel-level fusion methods. However, they were mainly designed based on a patch-based classification framework as shown in Fig. 1. For a given pixel G_α, a local square patch with a size of $t \times t$ is extracted and imported into the deep network for feature extraction and classification. For its neighborhood pixel G_β, the same process is conducted. Accordingly, patch-based classification methods are often time-consuming due to the fact that there are large amounts of overlaps between adjacent patches. In addition, a patch may break spatial integrity and connectivity of land-objects, which may lead to inaccurate classification results. Moreover, an optimal patch size is usually defined manually according to either a domain expert experience or multiple trials by considering object size and image resolution. This affects classification efficiency, especially when dealing with VHR MS and PAN images.

In addition to the patch-based classification methods, there are two kinds of patch-free classification methods: local patch-free classification (dense sample case) method and global patch-free classification (sparse sample case). Local patch-free classification (dense sample case) methods are also named as semantic segmentation in computer vision. The input has the same size of the image rather than that of patches generated for each pixel. They have been successfully applied in remote sensing image segmentation [24, 25]. As shown in Fig. 2, an encoder and decoder DL network is trained on a series of fixed-size (e.g., 512 × 512 pixels) local remote sensing images and the corresponding dense sample maps. Then, the trained network is used for predicting the semantic label of other images. However, this segmentation model requires large amounts of training data to learn the stable relationship between images and class labels, where dense training sample maps are

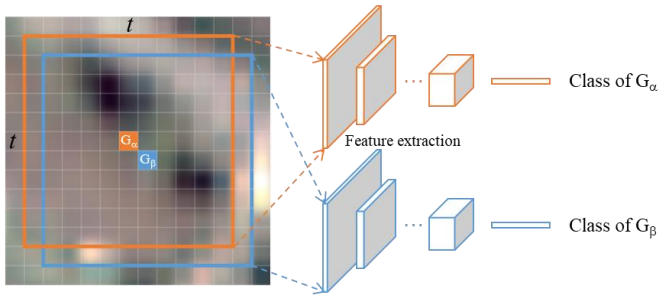


Fig. 1 Illustration of the patch-based classification scheme.

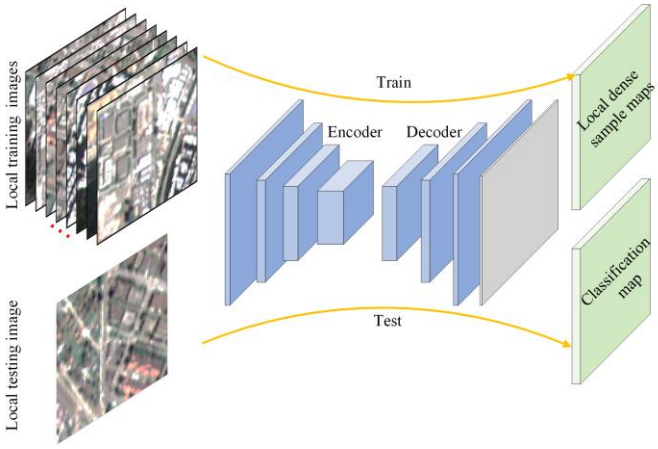


Fig. 2 Illustration of the patch-free classification scheme.

required [26]. Such requirements are difficult to meet in most of practical applications, since samples are usually manually selected from the image, the location of the samples is irregular and sparse, and the sample size is small. In [10], the global patch-free classification (sparse sample case) method was used, a group spatial-spectral attention fusion network was proposed for MS and PAN feature-level fusion and classification. Differently from the previous two classification methods, the global image is imported to the deep network to freely utilize the rich context information for feature extraction and classification in this global patch-free classification method, and the train sample can be sparse.

On the basis of this short analysis we identified some open issues that need to be further addressed: 1) *Problems related to the classification process*. The spatial integrity and connectivity of image objects may be affected by patch-based classification methods, and the overlapped patches between adjacent pixels inevitably increase the data processing burden. Moreover, an optimal patch size is usually estimated by a trial and error procedure to reach a balance between object size and image resolution. It is necessary to investigate how to build a patch-free network for MS and PAN image classification. 2) *Problems related to the fusion strategy*. Most of existing fusion methods only consider the relatively abundant cross-modal features. However, the homogeneous spectral information contained in MS features that is beneficial for identifying different objects may be destroyed by injecting spatial details of the PAN image. Also, the detailed spatial information contained in PAN features that can be used to describe edges of objects may be blurred after being fused with low-resolution

MS features. How to design a robust fusion process that captures the cross-modal features as well as spectral-spatial features from the two sources becomes a critical task.

To address the above open issues, in this work, a global collaborative fusion network (GCFnet) is proposed for MS and PAN image classification. Experimental results obtained on three real datasets confirm the effectiveness of the proposed GCFnet in terms of higher accuracy and robustness compared to the state-of-the-art methods. The main contributions of this paper are highlighted as follows.

1) A global patch-free classification scheme is proposed. It is designed to avoid the use of patches on all image pixels and thus can utilize the rich context information in the whole image to better model the spatial integrity and connectivity of image objects. Meanwhile, it can reduce the computational burden when compared to patch-based methods.

2) Unlike existing fusion methods that only use fused MS and PAN features for classification, a novel collaborative fusion architecture is proposed to take into account both shallow-to-deep features of MS and PAN images and their multiscale cross-modal features. In particular, an adaptively weighted loss function is proposed to calculate the total loss of individual and cross-modal branches of MS and PAN features, and a probability weighted decision-level fusion strategy is utilized to further improve the classification performance.

3) The proposed GCFnet utilizes both sampled and non-sampled position data in the training process. This allows to model land-objects in the image scene more comprehensively, and also meets the requirement of sufficient training data for a DL network. This is promising especially when dealing with a small sample set.

The rest of this paper is organized as follows. Related work is introduced in Section II. The proposed approach is described in detail in Section III. Experimental results and the related analysis are presented in Section IV. Finally, Section V draws the conclusions.

II. Proposed Feature-Level Fusion Approach

Aiming at effectively fusing the individual and cross-modal features of MS and PAN images, the proposed GCFnet consists in a novel collaborative fusion architecture with a global patch-free classification scheme. It mainly includes three fusion parts: 1) individual shallow-to-deep fusion, 2) multiscale cross-modal fusion, and 3) decision fusion and classification.

Fig. 3 shows the block diagram of the proposed GCFnet approach. It contains encoder and decoder modules, and is developed based on the global patch-free classification (sparse sample case) method. As an input, pixels of the global MS images (X_{MS}) and PAN images (X_{PAN}) are encoded to extract features, and then are decoded to revert to the original image size. Let $X_{MS} \in R^{B\times H\times W}$ and $X_{PAN} \in R^{1\times nH\times nW}$ be the input image, and the spatial resolution ratio between PAN and MS images be n. The predicted probability map $P \in R^{C\times H\times W}$ of each branch can be expressed as:

$$P = f(X_{MS}, X_{PAN}) \tag{1}$$

where f is the mapping model: $(R^{B\times H\times W}, R^{1\times nH\times nW}) \to R^{C\times H\times W}$ contains encoder and decoder, B is the number of input image bands, and C is the number of classes. During the training process, the global MS and PAN images are imported to the deep network for feature extraction and classification, whereas only the sparse training sample data is participated in the loss calculation (non-sampled position data is ignored). Then, the trained model is used for the same global MS and PAN images to predict the whole classification map. The global patch-free classification scheme does not rely on patches, where the training and predicting images are the same, thus features extracted during training can be directly used for prediction without network migration.

The encoder module mainly includes the convolutional block, the spectral attention block and the down-sampling layer, where the convolutional block contains the convolutional layer, the normalization and an activation function. The decoder module mainly consists of the convolutional layer and the upsampling layer. Unlike the existing methods that up-sample the MS image before feeding it into the network [8, 10], in our architecture, the original low-resolution MS image ($H\times W$) is directly fed into the network and its size is changed to $nH\times nW$ after encoding and decoding to avoid the bias and computational burden introduced by the up-sampling process. Note that in order to acquire the encoded features at the same size of the PAN image, the down-sampling ratio for the MS image is defined to be smaller than the PAN image.

A. Individual Shallow-to-deep Fusion

Two individual MS and PAN branches that contain encoding and decoding operations are defined, which aim to obtain the representative shallow-to-deep features from the original MS and PAN images. In the encoding stage, features are extracted from MS and PAN images. Then, the last MS and PAN encoded features with high-level semantic information are up-sampled into finer spatial resolution through successive up-sampling layers to recover the spatial resolution and achieve the pixel-based image classification result in the decoder architecture.

The shallower architecture has more spatial detail features, whereas the deeper architecture has more semantic features. In order to introduce the spatial details into semantic features for classification, the pointwise addition is used to fuse the shallow and deep features in each of the MS and the PAN branches. As shown in Fig. 3, a 1×1 convolutional layer is used after encoded feature to make the number of encoding channels the same as the deep decoding channels.

CNN is shown to be effective for tackling visual tasks and a series of convolutional layers in CNN focuses on spatial features of natural images [27, 28]. However, in remote sensing applications, spatial information, and spectral information should be fully exploited. Different channels in the feature maps usually represent different image objects [29], so the spectral attention mechanism is introduced for the MS image in the shallow encoding stage. The squeeze-and-excitation (SE) [30] blocks are used to adaptively recalibrate channel-wise

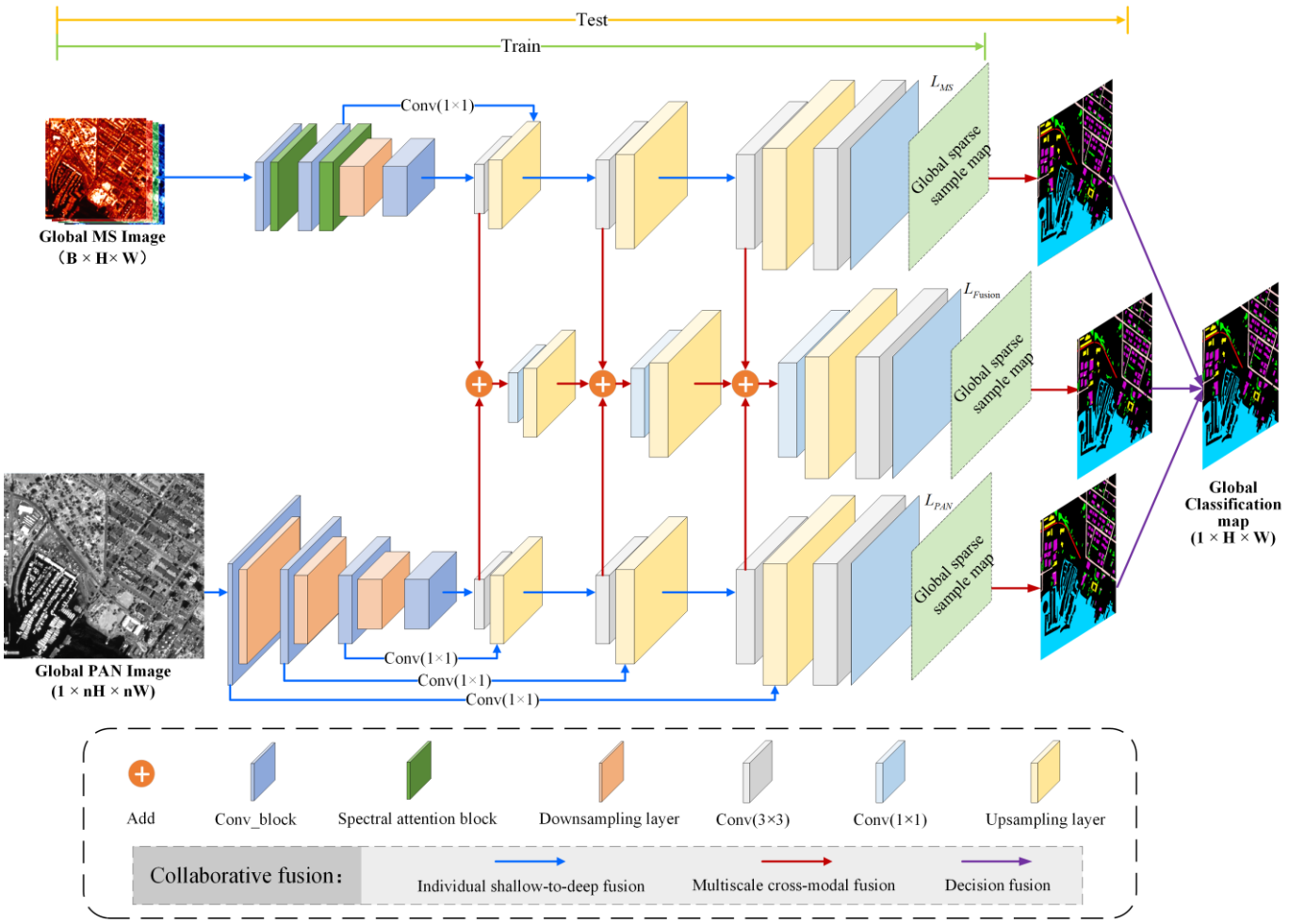


Fig. 3 Block diagram of the proposed GCFnet.

feature responses to emphasize important channels while suppressing noise.

As shown in Fig. 3, the global MS and PAN images are used as input, thus the batch size is always equal to one. In the case of a small batch size, the error of batch normalization (BN) increases rapidly due to inaccurate batch statistics estimation. In this work, BN is replaced by the group normalization (GN),which divides channels into groups and computes within each group the mean and variance values for normalization [31]. Instead of the commonly used maxpool layer, a convolutional layer with a stride of two is used for down-sampling, so that it can align the projected spatial location with its receptive field center leading to a more robust MS and PAN image classification result.

B. Multiscale Cross-modal Fusion

The cross-modal fusion branch is contained in the encoder module (see in Fig. 3). It includes a convolutional layer and an upsampling layer to fuse features of MS and PAN images. During the decoding stage feature resolution is gradually increased till reaching the same level of the classification map. In order to progressively fuse cross-modal features of MS and PAN images at different scales, the pointwise addition is used at each decoder scale. As shown in Fig. 3, the *i*-th layer decoded features of individual MS, PAN and cross-modal branches are denoted as F_{MS}^{i}, F_{PAN}^{i} and F_{Fusion}^{i} ($i = 1, 2, 3$), respectively. The multiscale cross-modal feature fusion can be formulated as follows:

$$F_{Fusion}^{i} = \begin{cases} F_{MS}^{i} + F_{PAN}^{i}, & i = 1 \\ F_{Fusion}^{i-1} + F_{MS}^{i} + F_{PAN}^{i}, & i = 2, 3 \end{cases} \tag{2}$$

The last convolutional layer of each branch that has *C* filters is used to perform the pixel-wise classification. The class probability $P_{(u,v),j}$ of a given pixel at location (u, v) in the *j*th category can be calculated by the softmax function:

$$P_{(u,v),j} = \frac{e^{g_{(u,v),j}}}{\sum_{k}^{C} e^{g_{(u,v),k}}} \tag{3}$$

where g denotes the last convolutional layer, $g_{(u,\ v),j}$ is the feature value in (u, v) of the *j*th channel of the last convolutional layer, and $\sum_{j=1}^{C} P_{(u,v),j} = 1$.

The cross-entropy loss is calculated to penalize the differences between the final output layer and the global reference map. It is defined as follows:

$$L = -\sum_{u=1}^{H}\sum_{v=1}^{W}\sum_{j=1}^{C} y_{(u,v),j} \log(P_{(u,v),j}) \tag{4}$$

where y is the reference map that is encoded in the form of one-hot. Let L_{MS}, L_{PAN} and L_{Fusion} be loss values of MS, PAN

and cross-modal fusion branch, respectively. In order to combine the three branches and achieve the optimal result, an adaptively weighted loss L_{total} can be formulated as:

$$L_{total} = \lambda_1 L_{MS} + \lambda_2 L_{PAN} + \lambda_3 L_{Fusion} \quad (5)$$

where λ_1, λ_2 and λ_3 represents the weights for L_{MS}, L_{PAN} and L_{Fusion}, respectively. Here λ_1, λ_2 and λ_3 are trainable parameters. Note that before network training, λ_1, λ_2 and λ_3 are all normalized to [0-1] by using the sigmoid function.

C. Decision Fusion and Classification

By following the network training and testing, three classification maps can be obtained by the three branches. Then a probabilistic weighted decision fusion scheme is adopted to fuse the three probabilistic maps. Let P_{MS}^j, P_{PAN}^j and P_{Fusion}^j (j=1, 2, ···, C) be the predicted probabilistic classification maps of MS, PAN and cross-modal fusion branches, respectively. The weighted probability decision-level fusion result of the j-th channel can be formulated as:

$$P_{DF}^j = \lambda_1 P_{MS}^j + \lambda_2 P_{PAN}^j + \lambda_3 P_{Fusion}^j, j = 1, \cdots, C \quad (6)$$

It is worth noting that λ_1, λ_2, and λ_3 are the final weights learned in the network according to Equation (5). The final probability classification map $P_{DF} \in R^{C \times H \times W}$ can be acquired by stacking weighting probabilistic maps of each class as follows:

$$P_{DF} = [P_{DF}^1, P_{DF}^2, \cdots, P_{DF}^C] \quad (7)$$

The final predicted classification map $\hat{y} \in R^{H \times W}$ can be obtained by assigning the pixel with the label having the maximum class probability, i.e.,

$$\hat{y} = \arg\max_C P_{DF} \quad (8)$$

The proposed GCFnet operations are summarized as follows.

Algorithm: Proposed GCFnet

Input: 1) the MS image X_{MS}; 2) the PAN image X_{PAN}; 3) the ground truth map y.

1: Input X_{MS} to the individual MS shallow-to-deep branch network and generate encoded and decoded features;

2: Input X_{PAN} to the individual PAN shallow-to-deep branch network and generate encoded and decoded features;

3: Fuse shallow and deep features in each individual MS and PAN branch by pointwise addition;

4: Multiscale cross-modal feature fusion between two data by Equation (2)

5: Produce probability classification maps P_{MS}, P_{PAN} and P_{Fusion} by Equation (3)

6: Calculate the adaptive weighting loss L_{total} by Equations (4) and (5), and update parameters from Steps 1 to 5 until L_{total} meets the requirement;

7: Calculate the decision fusion classification map by Equations (6), (7) and (8)

Output: Classification map

Table 1 Parameters of the three considered satellite images.

Satellites	MS image		PAN image	
	Resolution (*m*)	Band range (*μm*)	Resolution (*m*)	Band range (*μm*)
DEIMOS-2	4	B: 0.42-0.51 G: 0.51-0.58 R: 0.60-0.72 NIR:0.76-0.89	1	0.45-0.90
GaoFen-2	4	B:0.45-0.52 G: 0.52-0.59 R: 0.63-0.69 NIR: 0.77-9.89	1	0.45-0.90
QuickBird	2.4	B: 0.45-0.52 G: 0.52-0.66 R: 0.63-0.69 NIR: 0.76-0.90	0.6	0.45-0.90

Notes: B: Blue band; G: Green band; R: Red band; NIR: Near infrared band

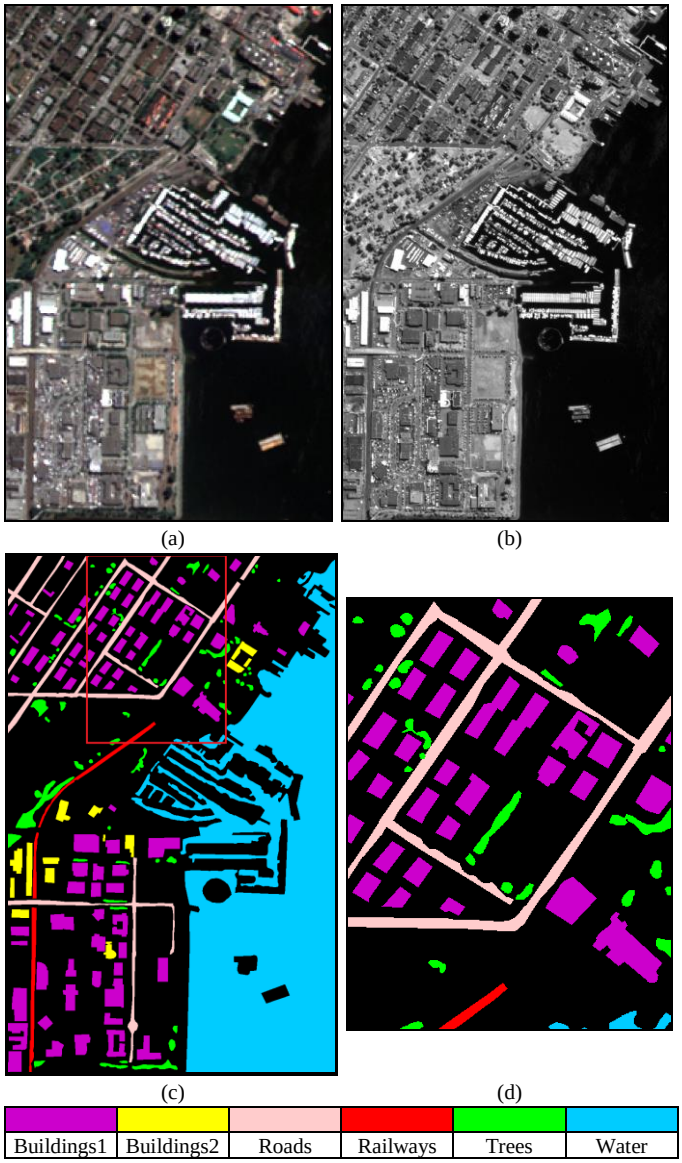


Fig. 4 VC dataset: (a) color composite of the MS image, (b) PAN image, (c) ground reference map, and (d) zoom of the portion of the image highlighted in the red box in (c).

Table 2 Numbers of training and testing samples in the VC dataset.

Classes		Number of samples (pixels)	
No.	**Name**	**Train**	**Test**
1	Buildings1	200	89867
2	Buildings2	200	13519
3	Roads	200	38206
4	Railways	200	9330
5	Trees	200	20703
6	Water	200	347554

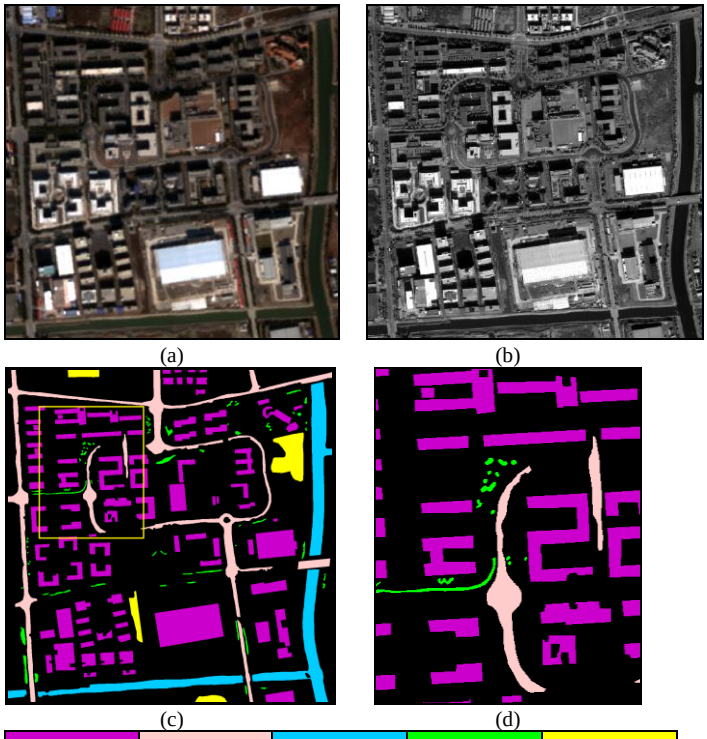


Fig. 5 SH dataset: (a) color composite of the MS image, (b) the PAN image, (c) the ground reference map, and (d) zoom of the portion of the image highlighted in the yellow box in (c).

Table 3 Numbers of training and testing samples in the SH dataset.

Classes		Number of samples (pixels)	
No.	**Name**	**Train**	**Test**
1	Buildings	200	195239
2	Roads	200	84244
3	Water	200	77843
4	Trees	200	11181
5	Grasses	200	24668

III. Experimental Results and Analysis

To evaluate the effectiveness of the proposed GCFnet, three multi-resolution VHR remote sensing datasets acquired by three satellites (i.e., DEIMOS-2, GaoFen-2 and QuickBird) were used in the experiments. Detailed parameter information is provided in Table 1. Note that reference maps of the three datasets were generated according to a careful image interpretation task. The spatial resolution of the reference maps is fixed the same as that of the corresponding PAN images.

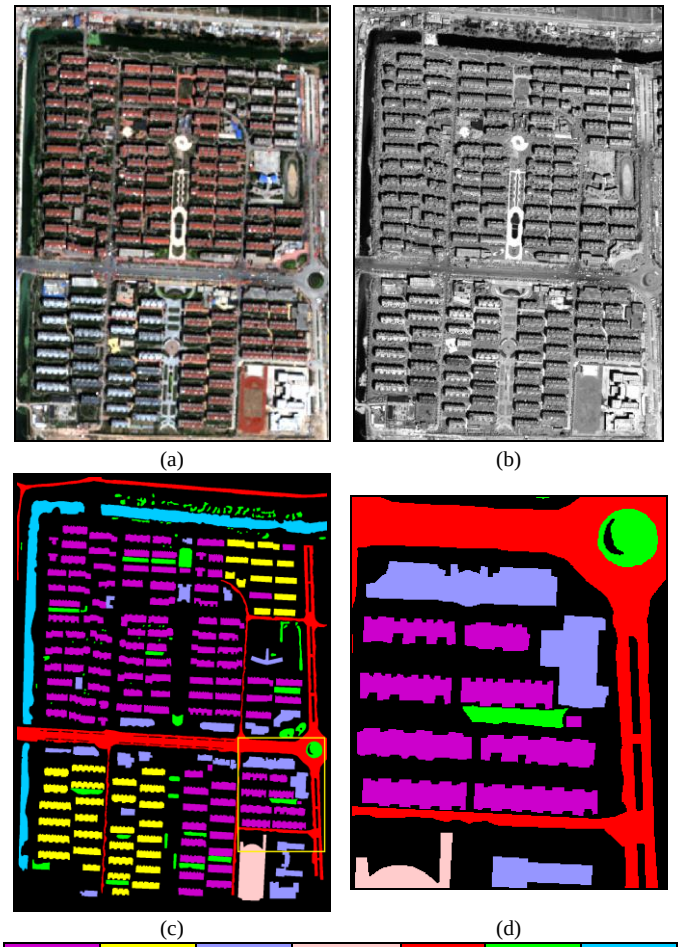


Fig. 6 XZ dataset: (a) true-color composite of the MS image, (b) PAN image, (c) ground reference map, and (d) zoom of the portion of the image highlighted in the yellow box in (c).

Table 4 Numbers of training and testing samples in the XZ dataset.

Classes		Number of samples (pixels)	
No.	**Name**	**Train**	**Test**
1	Buildings1	200	209250
2	Buildings2	200	80919
3	Buildings3	200	55554
4	Playground	200	18152
5	Roads	200	113484
6	Vegetation	200	41040
7	Water	200	72620

A. *Datasets Description*

1) Vancouver (VC) dataset: The MS and PAN images were acquired on March 31 and May 30, 2015, respectively, by the DEIMOS-2 satellite over Vancouver city, Canada. This dataset was provided by the 2016 IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest [32]. A subset of the whole image scene was cropped for our experiments. Spatial resolutions of the MS (with blue, green, red, and near-infrared four bands) and the PAN image are 4 m and 1 m, respectively. The sizes of the MS and PAN images are 345 × 219 and 1380 × 876 pixels, respectively. There are six classes

Table 5 Detailed parameters of the proposed GCFnet architecture.

		Individual MS branch		Cross-modal fusion branch		Individual PAN branch	
	Output size	Module	Parameters	Module	Parameters	Module	Parameters
Encoder	2H×2W	-	-	-	-	Conv_block Downsampling	[3×3]:32 [3×3]:64
	H×W	Conv_block SE block	[3×3]:64 r = 16	-	-	Conv_block Downsampling	[3×3]:64 [3×3]:128
	$\frac{H}{2}\times\frac{W}{2}$	Conv_block SE block Downsampling Conv_block	[3×3]:128 r = 16 [3×3]:256 [3×3]:256	-	-	Conv_block Downsampling Conv_block	[3×3]:128 [3×3]:256 [3×3]:256
Shallow-to-deep fusion	H×W	Conv	[1×1]:128	-	-	Conv	[1×1]:128
	2H×2W	-	-	-	-	Conv	[1×1]:128
	4H×4W	-	-	-	-	Conv	[1×1]:128
Decoder	$\frac{H}{2}\times\frac{W}{2}$	Conv	[3×3]:128	Conv	[1×1]:128	Conv	[3×3]:128
	H×W	Upsampling Conv	Interpolation (2) [3×3]:128	Upsampling Conv	Interpolation (2) [1×1]:128	Upsampling Conv	Interpolation (2) [3×3]:128
	2H×2W	Upsampling Conv	Interpolation (2) [3×3]:128	Upsampling Conv	Interpolation (2) [1×1]:128	Upsampling Conv	Interpolation (2) [3×3]:128
	4H×4W	Upsampling Conv Conv	Interpolation (2) [3×3]:128 [1×1]:C	Upsampling Conv Conv	Interpolation (2) [3×3]:128 [1×1]:C	Upsampling Conv Conv	Interpolation (2) [3×3]:128 [1×1]:C

in the scene, including buildings1 (with brown roofs), buildings2 (with white roofs), roads, railways, trees and water. Fig. 4(a) and (b) show the color composite of the MS image and its corresponding PAN image. Fig. 4 (c) shows the ground reference map, while the unlabeled area is in black. A portion of the image (highlighted in the red box in Fig. 4(c)) is zoomed in Fig. 4(d) for visual comparison. The numbers of training and testing samples are provided in Table 2. Based on the available reference map, 200 training samples were selected for each class and the remaining ones were used for testing.

2) Shanghai dataset (SH): This dataset was acquired on January 2, 2015, over Shanghai city, China, from Chinese GaoFen-2 satellite. The MS image is made up of 300 × 305 pixels with a spatial resolution of 4 m, while the PAN image has a size of 1200 × 1220 pixels with a spatial resolution of 1 m. Fig. 5 shows the color composite of the MS and PAN images, and the corresponding reference map. There are five land-cover classes (i.e., buildings, roads, water, trees and grasses) in the reference map. The numbers of training and testing samples used in the experiments are listed in Table 3.

3) Xuzhou dataset (XZ): The MS and PAN images in this dataset were acquired by the QuickBird satellite over the urban area of Xuzhou city, China. The spatial resolution of the MS and the PAN images are 2.4 m and 0.6 m, respectively. The sizes of the MS and PAN images are 283 × 379 and 1132 × 1516 pixels, respectively. Seven land-cover classes are presented in this scene, including buildings1 (with red roofs), buildings2 (with bluish roofs), buildings3 (with gray roofs), playground, roads, vegetation and water. The color composites of the two images and the corresponding reference maps are provided in Fig. 6. The numbers of training and testing samples used in the experiments are provided in Table 4.

B. Experimental Setup and Parameter Setting

Detailed parameters of the proposed GCFnet are listed in Table 5. Specifically, the kernel size of the convolution is 3×3 and the number of feature maps is 64. The reduction ratio *r* of the *SE* block is set as 16. The nearest neighbor interpolation with a factor of two is used to increase the size of MS image.

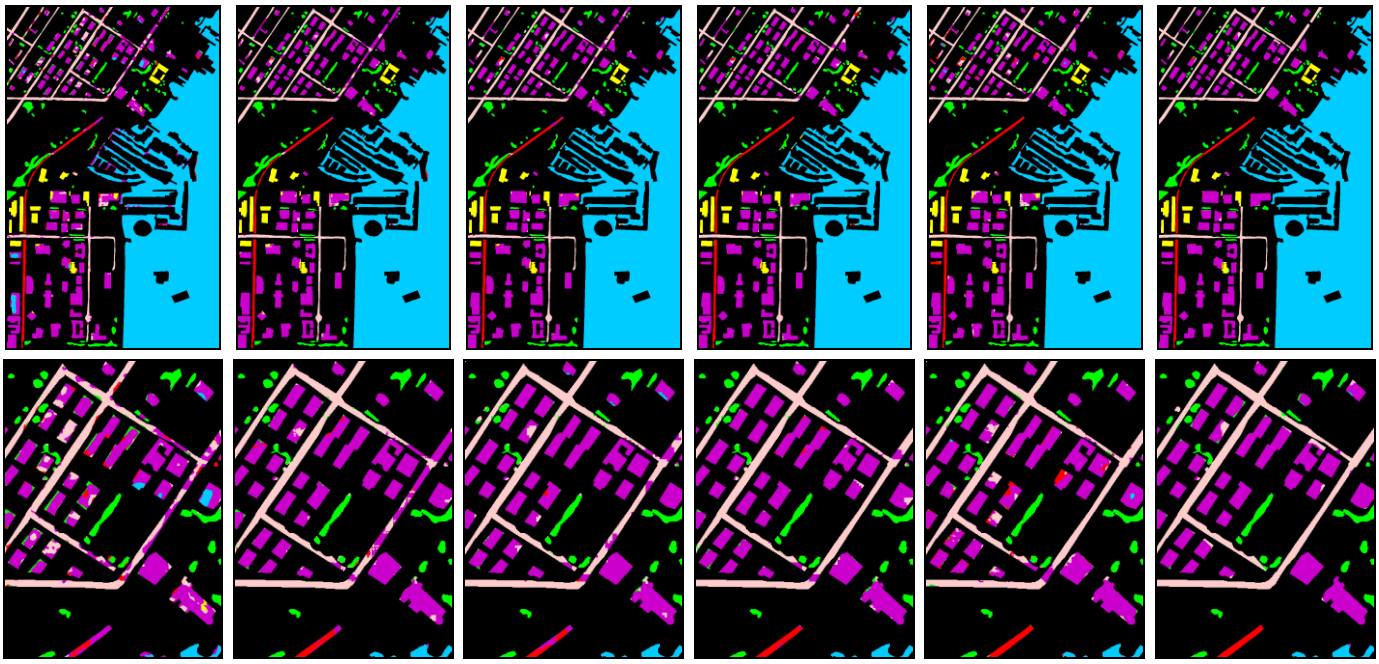
Based on multiple trials, the Adam optimization was used to set the learning rate as 0.0001. The number of training epochs was defined as 1000, and the batch size was set as 1.

Algorithms were implemented by using Python, where the DL networks were built by using Pytorch. Experiments were all carried out on the Ubuntu 18.04.5, with Intel(R) Xeon(R) Gold 6130 CPUs at 2.10GHz, 159-GB RAM, and GPU of NVIDIA GRID P40-24Q, 22GB.

C. Experimental Results

In order to validate the effectiveness of the proposed GCFnet in joint classification of MS and PAN images, four reference methods were considered for comparisons, including three state-of-the-art patch-based methods, i.e., the deep multiple instance learning (DMIL) [8], the MultiResolution Land Cover Classification (MultiResoLCC) [20], the cross-resolution hidden layer features fusion (CRHFF) [11], and the Group Attention Fusion network (GAFnet) in a global patch-free fusion [10]. It is important to note that final results were generated based on the average of ten times running of each method in order to test the robustness of the considered methods. Abbreviations F_{MS*PAN} and $F_{EMAP*PAN}$ represent the feature-level fusion based on the original MS and PAN images and based on the extended multi-attribute profile (EMAP) features and PAN image in the CRHFF approach, respectively. To compare and evaluate the obtained results quantitatively, the overall accuracy (OA), the average accuracy (AA), the Kappa coefficient (Kappa) and the per-class accuracies were calculated. The standard deviation (SD) value is also provided to further validate the stability of different methods.

1) Results on the VC dataset: Classification accuracies achieved by different methods on the VC dataset are listed in Table 6. Considering the two global patch-free methods, the proposed GCFnet achieved the highest accuracies (i.e., OA: 99.04%, AA: 98.59%, Kappa: 98.12%), outperforming the



(a) 94.56% (b) 97.68% (c) 98.33% (d) 98.74% (e) 97.64% **(f) 99.21%**

Fig. 7 Classification maps obtained by different methods on the VC dataset: (a) DMIL, (b) MultiResoLCC, (c) CRHFF (F_{MS*PAN}), (d) CRHFF ($F_{EMAP*PAN}$), (e) GAFnet, and (f) proposed GCFnet. The first row represents the whole classification maps at global scale, and the second row represents the subsets at local scale.

Table 6 Comparison of the classification accuracies (%) provided by different methods (VC dataset)

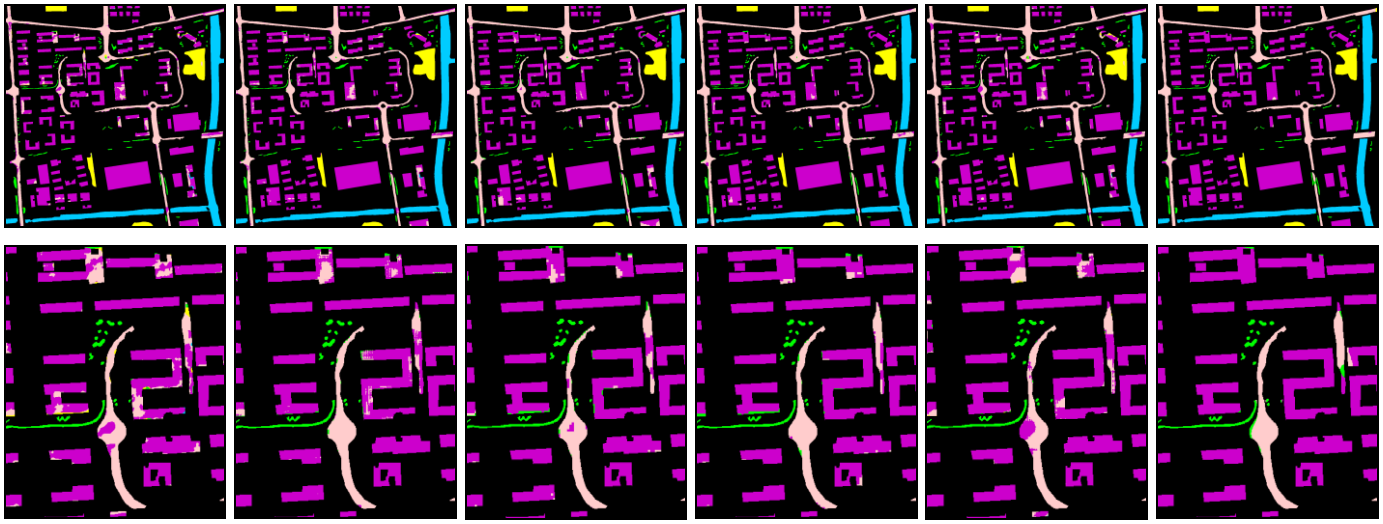
Class	Patch-based				Global patch-free	
	DMIL	MultiResoLCC	CRHFF		GAFnet	Proposed GCFnet
			F_{MS*PAN}	$F_{EMAP*PAN}$		
Buildings1	82.27±1.62	93.08±1.63	95.46±0.59	95.79±0.71	84.74±4.29	96.26±0.42
Buildings2	98.07±0.46	99.53±0.22	99.26±0.23	99.70±0.16	98.65±0.28	99.62±0.14
Roads	84.65±2.28	86.06±2.44	90.21±0.90	93.99±0.58	90.78±3.59	97.42±0.41
Railways	92.91±1.65	95.41±1.55	94.63±1.49	99.35±0.28	97.60±0.76	99.92±0.12
Trees	96.41±0.39	97.15±0.42	97.53±0.48	97.72±0.31	99.04±0.17	98.37±0.27
Water	97.79±0.53	99.70±0.16	99.87±0.02	99.80±0.03	99.72±0.09	99.93±0.03
OA	94.00±0.37	97.37±0.23	98.19±0.07	98.59±0.09	96.37±1.01	**99.04±0.08**
AA	92.02±6.92	95.16±5.11	96.16±3.56	97.72±2.39	95.09±1.40	**98.59±0.11**
Kappa	88.48±0.67	94.89±0.44	96.48±0.14	97.26±0.18	92.99±1.95	**98.12±0.15**

GAFnet method by increasing the OA, AA and Kappa of 2.67%, 3.5% and 5.13%, respectively. For most of the classes, the proposed GCFnet resulted in higher accuracies than the GAFnet. Class accuracies on buildings1 and roads were sharply increased by 11.52% and 6.64%, respectively. The SD of OA value of the proposed GCFnet (i.e., 0.08) is significantly smaller than the one of the GAFnet (i.e., 0.93), which indicates more stable classification performance.

Furthermore, the CRHFF($F_{EMAP*PAN}$) method achieved the best performance among all compared patch-based methods. It also outperformed the patch-free GAFnet in terms of higher accuracy and smaller SD values. The proposed GCFnet is superior to all patch-based methods by increasing the OA values of 0.45%, 0.85%, 1.67% and 5.04%, with respect to CRHFF($F_{EMAP*PAN}$), CRHFF(F_{MS*PAN}), MultiResoLCC and DMIL methods, respectively. The proposed GCFnet improves the OA of 0.45% over the CRHFF($F_{EMAP*PAN}$) method. It should be noted that the EMAP features were used (and thus some additional shallow spectral-spatial information were considered) in the CRHFF($F_{EMAP*PAN}$), whereas the proposed GCFnet method only uses the two original images. Some classes such as buildings1 and roads exhibited low accuracies in the reference methods due to their similar spectral representations. However, the proposed GCFnet still yielded the best performance on the two classes with accuracies equal to 96.26% (buildings1) and 97.42% (roads).

Classification maps are shown in Fig. 7. The global and local classification maps are shown in the first and second rows, respectively. One can see that most methods obtain good visual classification results for trees and water classes. However, false alarms are present between buildings1, roads and railways in the patch-based DMIL, MultiResoLCC and CRHFF (F_{MS*PAN}) methods and patch-free GAFnet. By adding EMAP features, the CRHFF($F_{EMAP*PAN}$) approach improved the classification results continuity inside those three classes. Furthermore, the GCFnet produced a better classification map than the CRHFF($F_{EMAP*PAN}$) in terms of maintaining the inner-homogeneity inside buildings1 and roads.

2) Results on the SH dataset: Table 7 provides the classification accuracies obtained by all compared methods on the SH dataset. The proposed GCFnet achieved the best performance among all methods. For the two global patch-free



(a) 93.74% (b) 95.80% (c) 96.72% (d) 98.24% (e) 95.74% **(f) 99.32%**

Fig. 8 Classification maps obtained by different methods on the SH dataset: (a) DMIL, (b) MultiResoLCC, (c) CRHFF (F_{MS*PAN}), (d) CRHFF ($F_{EMAP*PAN}$), (e) GAFnet, and (f) proposed GCFnet. The first row represents the whole classification maps at global scale, and the second row represents the subsets at local scale.

Table 7 Comparison of the classification accuracies (%) provided by different methods (SH dataset)

	Patch-based				Global patch-free	
Class	**DMIL**	**MultiResoLCC**	**CRHFF**		**GAFnet**	**Proposed GCFnet**
			F_{MS*PAN}	**$F_{EMAP*PAN}$**		
Buildings	89.59±1.23	93.96±0.50	94.99±0.36	98.00±0.09	92.21±2.53	98.82±0.37
Roads	89.56±1.04	93.60±0.61	95.29±0.34	95.92±0.35	91.94±1.47	98.25±0.36
Water	99.83±0.06	99.99±0.03	100.00±0.00	99.94±0.04	99.93±0.05	99.73±0.20
Trees	98.86±0.27	99.72±0.21	99.90±0.04	99.91±0.04	99.65±0.15	99.64±0.24
Grasses	97.38±1.05	99.73±0.41	99.98±0.04	99.99±0.03	99.20±0.43	99.92±0.10
OA	92.37±0.79	95.60±0.21	96.50±0.13	98.12±0.08	94.33±1.49	**98.97±0.20**
AA	95.04±5.07	97.40±3.31	98.03±2.64	98.75±1.79	96.59±0.71	**99.27±0.10**
Kappa	88.66±1.15	93.43±0.31	94.77±0.19	97.17±0.13	91.55±2.18	**98.45±0.29**

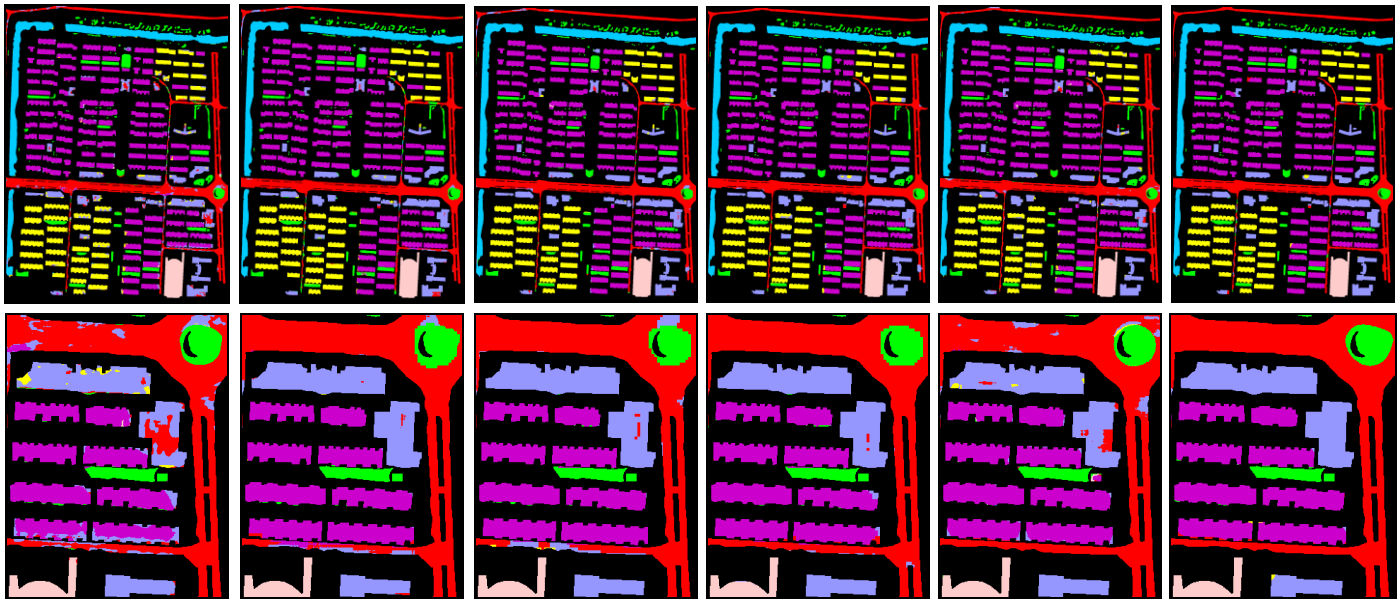
methods, the average OA, AA and Kappa values of the GCFnet are 98.97%, 99.27% and 98.45%, which are 4.64%, 2.68%, and 6.9% higher than those of the GAFnet. In particular, significant improvements were achieved by the GCFnet on buildings and roads classes, yielding an increase of the accuracy of 6.61% and 6.31%, respectively. On the other hand, the proposed GCFnet also outperformed all patch-based methods in terms of OA, AA and Kappa values. In particular, it yielded higher class accuracies than CRHFF($F_{EMAP*PAN}$) on the two easily confused classes, i.e., buildings and roads, resulting in improvements of 0.82% (from 98.00% to 98.82%) and 2.33% (from 95.92% to 98.25%), respectively. The CRHFF($F_{EMAP\ *\ PAN}$) method outperformed the other patch-based methods and the patch-free GCFnet with higher OA but lower SD values.

From the classification maps shown in Fig. 8, one can see that the proposed GCFnet produced the best overall classification results at global scale (see Fig. 8). To better compare the visual classification results for buildings and roads classes, local regions that mainly contain buildings and roads are extracted and compared in the second row of Fig. 8. It is obvious that the proposed GCFnet produced a better inner-homogeneity and finer boundaries for buildings and roads classes when compared to the reference methods.

3) Results on the XZ dataset: Table 8 lists the quantitative results for the XZ dataset. The proposed GCFnet achieved the highest accuracies (i.e., OA = 98.59%, AA = 98.71%, Kappa = 98.22%) among all compared methods. The OA value of the proposed GCFnet is 4.24% higher than that of the GAFnet, with improvements mainly focused on classes of buildings1, buildings2, building3 and roads. The CRHFF($F_{EMAP\ *\ PAN}$) yielded the highest accuracies among all patch-based methods. It benefited from using the shallow EMAP features. The OA value obtained by the CRHFF($F_{EMAP*PAN}$) method is 0.88% higher than to the one obtained by the CRHFF(F_{MS*PAN}) method. However, the latter still outperformed DMIL and MultiResoLCC methods, with improvements in OA values equal to 3.8% and 0.95%, respectively. The SD of OA values in the proposed GCFnet method is 0.08, which is the smallest among all methods.

Fig. 9 shows classification maps obtained at global and local scales. Compared with other datasets, the XZ dataset contains more complex classes (e.g., buildings1, buildings2, buildings3 and roads). From the local classification map (Fig. 9 row 2), one can see that the CRHFF($F_{EMAP*PAN}$) method obtained better visual classification result than other reference methods. However, it still contain some noise, especially in the interior and boundaries of buildings. On the contrary, the proposed GCFnet outperforms CRHFF($F_{EMAP*PAN}$) with smoother and more complete maps also on such complex classes.

Fig. 10 shows the SD values of the OA obtained by different methods on the three datasets. It is clear that in all three datasets,



(a) 94.31% (b) 97.26% (c) 97.60% (d) 98.40% (e) 97.50% **(f) 98.71%**

Fig. 9 Classification maps obtained by different methods on the XZ dataset: (a) DMIL, (b) MultiResoLCC, (c) CRHFF (F_{MS*PAN}), (d) CRHFF ($F_{EMAP*PAN}$), (e) GAFnet, and (f) proposed GCFnet. The first row represents the whole classification maps at global scale, and the second row represents the subsets at local scale.

Table 8 Comparison of the classification accuracies (%) provided by different methods (XZ dataset)

Class	Patch-based				Global patch-free	
	DMIL	MultiResoLCC	CRHFF		GAFnet	Proposed GCFnet
			F_{MS*PAN}	$F_{EMAP*PAN}$		
Buildings1	95.43±0.72	98.78±0.35	98.69±0.14	98.68±0.25	94.85±2.11	98.88±0.13
Buildings2	90.98±1.46	96.69±4.80	98.79±0.17	98.50±0.42	92.37±5.81	99.50±0.15
Buildings3	84.22±1.31	95.75±1.83	96.47±0.78	97.08±0.82	88.05±9.34	98.10±0.59
Playground	99.89±0.07	99.97±0.06	99.98±0.07	99.91±0.03	99.76±0.41	100.00±0.00
Roads	90.68±2.70	89.66±5.48	92.98±0.88	96.95±0.62	92.16±5.99	96.92±0.34
Vegetation	97.71±0.35	96.80±0.67	96.73±0.34	97.88±0.22	98.36±0.16	97.74±0.45
Water	99.07±0.26	99.58±0.30	99.55±0.04	99.73±0.14	99.71±0.17	99.85±0.06
OA	93.60±0.81	96.45±1.68	97.40±0.16	98.28±0.14	94.35±3.34	**98.59±0.08**
AA	94.00±5.65	96.71±3.50	97.57±2.38	98.39±1.23	95.04±3.19	**98.71±0.12**
Kappa	91.93±1.01	95.52±2.10	96.72±0.21	97.83±0.18	92.88±4.19	**98.22±0.11**

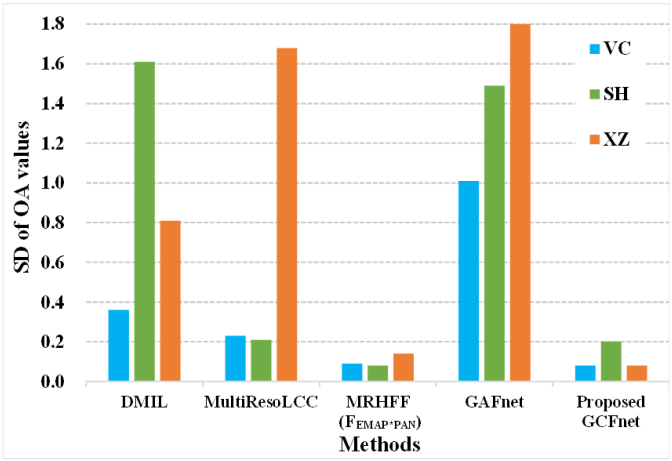
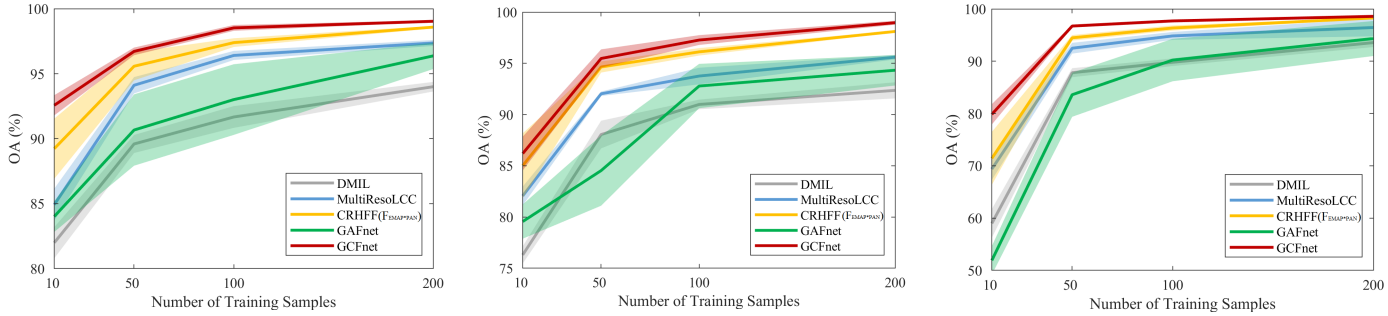


Fig. 10 Comparison of the SD of OA values obtained by different methods on the three considered datasets.

the SD of OA values in the proposed GCFnet is close to that of CRHFF($F_{EMAP*PAN}$), which is much lower than those of the compared state-of-the-art methods. This further demonstrates the robustness of the proposed GCFnet.

4) Comparison of computing time: Table 9 shows the training and testing time of different methods. The training time is related to the number of training samples (200 samples for each class were selected in our case), as well as the architecture and the number of parameters. One can be seen that the training process was very costly in the CRHFF method, since hidden layer features were extracted additionally. For the testing process, the two global patch-free methods were much faster than the patch-based methods (i.e., DMIL, MultiResoLCC and CRHFF), which inevitably increased the computational complexity due to processing of overlapped patches. However, the global patch-free methods (i.e., GAFnet and the proposed GCFnet) performed in a more efficient way.

5) Comparison of results obtained by different sample sizes: Fig. 11 shows the OA values obtained by different methods with different number of training samples (i.e., 10, 50, 100, 200 samples), where each curve represents the average OA value and the shaded area represents the SD of OA after running ten times. The proposed GCFnet yielded the best performance regardless of the sample size, even in the extremely small sample-size case when only 10 training samples were used. Taking the VC dataset as an example, when 200 samples were considered, the proposed GCFnet achieved the highest average OA = 99.04%, outperforming the DMIL method by 5.04%.



(a) (b) (c)

Fig. 11 Comparison of the OA values obtained by different methods with different number of training samples on the three considered datasets: (a) VC dataset; (b) SH dataset; (c) XZ dataset.

Table 9 Computing time (h: hours, s: seconds) of the training and the testing processes for different methods

Classification Fashion	Methods	Processes	Datasets		
			VC	SH	XZ
Patch-based	DMIL	Training(h)	0.24	0.32	0.57
		Testing(s)	216	283	502
	MultiResoLCC	Training(h)	0.10	0.10	0.20
		Testing(s)	143	109	291
	CRHFF ($F_{EMAP*PAN}$)	Training(h)	3.44	4.16	4.88
		Testing(s)	121	93	253
Patch-free	GAFnet	Training(h)	0.26	0.32	0.37
		Testing(s)	0.31	0.35	0.49
	Proposed GCFnet	Training(h)	0.17	0.19	0.24
		Testing(s)	0.31	0.34	0.51

Table 10 Comparison of the classification accuracies (%) provided by the cross-modal fusion only and the whole collaborative fusion (VC dataset)

Class	Only cross-modal fusion	Collaborative fusion
Buildings1	94.19±0.48	96.55±0.30
Buildings2	99.49±0.33	99.78±0.09
Roads	97.52±0.45	97.86±0.22
Railways	99.58±0.25	99.61±0.51
Trees	98.37±0.21	98.74±0.17
Water	99.84±0.08	99.86±0.12
OA	98.62±0.10	99.09±0.09
AA	98.17±0.14	98.73±0.11
Kappa	97.32±0.19	98.23±0.18

Table 11 Comparison of the classification accuracies (%) provided by the individual branches and the one obtained after decision fusion step (VC dataset)

Class	MS Branch	PAN Branch	Cross-modal Fusion	Decision Fusion
Buildings1	94.65±0.40	90.46±0.91	96.22±0.35	96.55±0.30
Buildings2	99.44±0.18	99.80±0.09	99.76±0.15	99.78±0.09
Roads	96.48±0.66	96.68±0.54	97.85±0.23	97.86±0.22
Railways	99.13±0.73	99.43±0.47	99.66±0.53	99.61±0.51
Trees	98.28±0.24	97.41±0.46	98.69±0.19	98.74±0.17
Water	99.69±0.18	99.85±0.05	99.87±0.13	99.86±0.12
OA	98.51±0.18	97.88±0.13	99.04±0.10	99.09±0.09
AA	97.95±0.23	97.27±0.21	98.67±0.11	98.73±0.11
Kappa	97.12±0.35	95.90±0.26	98.13±0.18	98.23±0.18

Furthermore, the proposed GCFnet obtained the highest OA = 92.57% in the case of only 10 samples, exceeding by 10.58% the one in the DMIL method. In the extremely small training set, the difference between the proposed GCFnet and DMIL was larger than in other cases. Similar results can be also obtained in the other two considered datasets as shown in Fig. 11. This confirms the advantage and effectiveness of the proposed GCFnet method in dealing with the challenging small sample size issue.

Fig. 11 shows the accuracies obtained by all methods increased as the sample size increased, and become stable when the sample size exceeded 100. It is worth noting that that for most methods, the SD values decreased by increasing the sample size. The proposed GCFnet and MultiResoLCC resulted in more stable performance among all methods, and the CRHFF($F_{EMAP*PAN}$) performed better when the sample size exceeded 100.

6) Analysis of Collaborative Fusion: To analyze the effectiveness of collaborative fusion in the proposed GCFnet, we compared classification results obtained by both the cross-modal fusion only and the whole collaborative fusion on the VC dataset. As shown in Table 10, the whole collaborative fusion is superior to the case using only the cross-modal fusion component in terms of higher OA values and lower SD values. The final learned weights λ_1, λ_2 and λ_3 of L_{MS}, L_{PAN} and L_{Fusion}, are 0.0552, 0.0314 and 0.0619, respectively. The L_{Fusion} highest weight value demonstrates that the cross-modal fusion features play a more important role than the two individual features in classification. However, the integration of individual MS and PAN features can further enhance the feature representation thus leading to an improved classification result.

To evaluate the performance of the decision fusion step in the collaborative fusion, classification accuracies obtained by the individual MS, PAN and the cross-modal fusion branches were compared with the one obtained by the decision fusion step (i.e., the final collaborative fusion result) in Table 11. It should be noted that the cross-modal fusion in Table 11 is calculated based on the proposed collaborative fusion method, thus is different from the one in Table 10, where only the cross-modal fusion classification map was calculated. From Table 11, one can observe that the cross-modal fusion branch achieved better accuracy than other two individual MS and PAN branches. By taking advantage of the decision fusion step, the classification performance is further enhanced in terms of higher OA values and lower SD values.

IV. Conclusion

In this paper, a global collaborative fusion network (GCFnet) has been proposed for joint classification of MS and PAN images. To avoid the use of overlapped patches and maintain the spatial integrity and connectivity, a global patch-free

classification scheme based on an encoder-decoder DL network was developed to exploit context dependencies in the image and improve classification efficiency. A novel collaborative fusion architecture was proposed to fuse shallow-to-deep features and multiscale cross-modal features extracted from MS and PAN images, where an adaptive loss weighted fusion strategy was designed to calculate the total loss of branches. In addition, a probability weighted decision fusion strategy was applied to fuse classification results of the three branches to further improve the classification output. Experimental results obtained on three real remote sensing datasets acquired by DEIMOS-2, GaoFen-2, and QuickBird satellites confirmed the effectiveness of the proposed approach. By comparing with the state-of-the-art methods, the proposed GCFnet resulted in higher classification accuracy, and more stable and efficient performances.

ACKNOWLEDGMENTS

The authors would like to thank Deimos Imaging for acquiring and providing the VC dataset, and the IEEE GRSS Image Analysis and Data Fusion Technical Committee. We also thank the China Centre for Resources Satellite Data and Application for providing the GaoFen-2 images.

REFERENCES

[1] S. Liu, D. Marinelli, L. Bruzzone, and F. Bovolo, "A Review of Change Detection in Multitemporal Hyperspectral Images: Current Techniques, Applications, and Challenges," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 2, pp. 140-158, 2019.

[2] S. Liu, Y. Zheng, M. Dalponte, and X. Tong, "A novel fire index-based burned area change detection approach using Landsat-8 OLI data," *European Journal of Remote Sensing*, vol. 53, no. 1, pp. 104-112, 2020.

[3] Y. Zheng, S. Liu, Q. Du, H. Zhao, X. Tong, and M. Dalponte, "A Novel Multitemporal Deep Fusion Network (MDFN) for Short-term Multitemporal HR Images Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 10691-10704, 2021.

[4] R. L. Hang, Z. Li, P. Ghamisi, D. F. Hong, G. Y. Xia, and Q. S. Liu, "Classification of Hyperspectral and LiDAR Data Using Coupled CNNs," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 7, pp. 4939-4950, Jul 2020.

[5] M. Zhang, W. Li, R. Tao, H. Li, and Q. Du, "Information Fusion for Classification of Hyperspectral and LiDAR Data Using IP-CNN," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-12, 2022.

[6] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, "More Diverse Means Better: Multimodal Deep Learning Meets Remote-Sensing Imagery Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4340-4354, 2021.

[7] W. Kang, Y. Xiang, F. Wang, and H. You, "CFNet: A Cross Fusion Network for Joint Land Cover Classification Using Optical and SAR Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 1562-1574, 2022.

[8] X. Liu, L. Jiao, J. Zhao, J. Zhao, D. Zhang, F. Liu, S. Yang, and X. Tang, "Deep Multiple Instance Learning-Based Spatial-Spectral Classification for PAN and MS Imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 1, pp. 461-473, Jan 2018.

[9] J. R. Bergado, C. Persello, and A. Stein, "Recurrent Multiresolution Convolutional Networks for VHR Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 11, pp. 6361-6374, 2018.

[10] X. Liu, L. Li, F. Liu, B. Hou, S. Yang, and L. Jiao, "GAFnet: Group Attention Fusion Network for PAN and MS Image High-Resolution Classification," *IEEE Transactions on Cybernetics*, vol. 51, pp 1-14, Mar 30 2021.

[11] S. Liu, H. Zhao, Q. Du, L. Bruzzone, A. Samat, and X. Tong, "Novel Cross-Resolution Feature-Level Fusion for Joint Classification of Multispectral and Panchromatic Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 2021.

[12] P. Du, S. Liu, J. Xia, and Y. Zhao, "Information fusion techniques for change detection from multi-temporal remote sensing images," *Information Fusion*, vol. 14, no. 1, pp. 19-27, 2013.

[13] S. Liu, Y. Zheng, Q. Du, A. Samat, X. Tong, and M. Dalponte, "A Novel Feature Fusion Approach for VHR Remote Sensing Image Classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 464-473, 2021.

[14] S. Liu, Y. Zheng, Q. Du, L. Bruzzone, A. Samat, X. Tong, Y. Jin, and C. Wang, "A Shallow-to-Deep Feature Fusion Network for VHR Remote Sensing Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-13, 2022.

[15] F. Palsson, J. R. Sveinsson, J. A. Benediktsson, and H. Aanaes, "Classification of Pansharpened Urban Satellite Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, no. 1, pp. 281-297, 2012.

[16] N. Kosaka, T. Akiyama, Bien Tsai, and T. Kojima, "Forest type classification using data fusion of multispectral and panchromatic high-resolution satellite imageries," *in Proceedings. IEEE International Geoscience and Remote Sensing Symposium*, Jul. 25-29, 2005, vol. 4, pp. 2980-2983.

[17] G. Moser, A. De Giorgi, and S. B. Serpico, "Multiresolution Supervised Classification of Panchromatic and Multispectral Images by Markov Random Fields and Graph Cuts," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 9, pp. 5054-5070, 2016.

[18] T. Mao, H. Tang, J. Wu, W. Jiang, S. He, and Y. Shu, "A Generalized Metaphor of Chinese Restaurant Franchise to Fusing Both Panchromatic and Multispectral Images for Unsupervised Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 8, pp. 4594-4604, 2016.

[19] W. Zhao, L. Jiao, W. Ma, J. Zhao, J. Zhao, H. Liu, X. Cao, and S. Yang, "Superpixel-Based Multiple Local CNN for Panchromatic and Multispectral Image Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 4141-4156, 2017.

[20] R. Gaetano, D. Ienco, K. Ose, and R. Cresson," A Two-Branch CNN Architecture for Land Cover Classification of PAN and MS Imagery," *Remote Sensing*, vol. 10, no. 11, 2018.

[21] J. R. Bergado, C. Persello, and A. Stein, "Fusenet: End-to-End Multispectral VHR Image Fusion and Classification," in Proceedings. *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 2091-2094, 2018.

[22] H. Zhu, W. P. Ma, L. L. Li, L. C. Jiao, S. Y. Yang, and B. Hou, "A Dual-Branch Attention fusion deep network for multiresolution remote-Sensing image classification," *Information Fusion*, vol. 58, pp. 116-131, Jun 2020.

[23] W. Ma, J. Shen, H. Zhu, J. Zhang, J. Zhao, B. Hou and L. Jiao, "A Novel Adaptive Hybrid Fusion Network for Multiresolution Remote Sensing Images Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1-17, 2021.

[24] H. Zhu, R. Tan, L. Han, H. Fan, Z. Wang, B. Du, S. Liu and Q. Liu, "DSSM: A Deep Neural Network with Spectrum Separable Module for Multi-Spectral Remote Sensing Image Segmentation," *Remote Sensing*, vol. 14, no. 4, 2022.

[25] D. Hong, J. Yao, D. Meng, Z. Xu, and J. Chanussot, "Multimodal GANs: Toward Crossmodal Hyperspectral–Multispectral Image Segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 6, pp. 5103-5113, 2020.

[26] S. Hong, S. Kwak, and B. Han, "Weakly Supervised Learning with Deep Convolutional Neural Networks for Semantic Segmentation: Understanding Semantic Layout of Images with Minimum Human Supervision," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 39-49, 2017.

[27] Jonathan Long, Evan Shelhamer, Trevor Darrell, "Fully convolutional networks for semantic segmentation," *in Proceedings. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431-3440, 2015*

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84-90, 2017.

[29] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, T. S. Chua, "SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning," *in Proceedings. IEEE Conference on*

Computer Vision and Pattern Recognition (CVPR), pp. 5659-5667, 2017.

[30] J. Hu, L. Shen and G. Sun, "Squeeze-and-excitation networks," *in Proceedings. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7132-7141, 2018.

[31] Y. Wu and K. He, "Group Normalization," *in Proceedings. European Conference on Computer Vision (ECCV)*, pp. 3-19, 2018.

[32] 2016 IEEE GRSS Data fusion contest. online: http://www.grss–ieee.org/community/technical–committees/data–fusion.