CFNet: A Cross Fusion Network for Joint Land Cover Classification Using Optical and SAR Images

Wenchao Kang, Yuming Xiang, *Member, IEEE*, Feng Wang, and Hongjian You

***Abstract*—As two of the most widely used remote sensing images, optical and synthetic aperture radar (SAR) images show abundant and complementary information on the same target owing to their individual imaging mechanisms. Consequently, using optical and SAR images simultaneously can better describe the inherent features of the target, and thus, be beneficial for subsequent remote sensing applications. In this article, we propose a novel modular fully convolutional network model to improve the accuracy of land cover classification by fully exploiting the complementary features of the two sensors. We investigate where and how to fuse the two images in the joint classification network. A cross-gate module with a bidirectional information flow is proposed to achieve the best fusion performance. In addition, to validate the proposed model, we construct a multiclass land cover classification dataset. Exhaustive experiments show that the proposed joint classification network presents superior results than state-of-the-art classification models using single-sensor images.**

***Index Terms*—Fully convolutional network (FCN), joint land cover classification, optical remote sensing image, synthetic aperture radar (SAR).**

I. Introduction

HYPERSPECTRAL images, multispectral images, LiDAR images (digital surface model, DSM), and synthetic aperture radar (SAR) images are the most commonly used remote sensing images. They can capture different features of the same ground object. Hyperspectral images can provide finer spectral information for ground object description and are particularly suitable for distinguishing between camouflaged targets with similar textures and different spectra. Hyperspectral images require spectral unmixing, but the spectrum of ground objects is disturbed by many factors and suffers from spectral variability [1]. The multispectral images reflect the color and brightness information of the target and have a high discriminative ability for city streets, buildings, water, soil, and vegetation. Optical images are interfered with by clouds, rain, and snow. SAR images mainly reflect two types of properties of the ground target: structural properties (texture, geometry, etc.) and electromagnetic scattering properties (dielectric properties, polarization properties). SAR images suffer from severe speckle noise and have special phenomena such as shadow and foreshortening. Theoretically, employing multisource images can exploit complementary information to improve the land cover classification accuracy. In addition, the use of multisource images can also compensate for the respective problems of single-source images. The use of multisource remote sensing imagery for land cover classification is, therefore, of research interest.

As the acquisition of multisource images has become easier, related studies have received more attention, especially after introducing deep learning. Both Volpi *et al.* [2] and Audebert *et al.* [3] used multispectral images and DSM data to implement land cover classification. In [2], a neural network was only used to extract features, while classification was performed by an extra traditional classifier. Audebert *et al.* [3] instead designed fully convolutional networks (FCN) to perform the two tasks at the same time. The winner of the 2018 IEEE Data Fusion Competition further proposed a three-branch FCN to simultaneously utilize hyperspectral images, multispectral images, and LiDAR images [4]. Besides, Srivastava *et al.* [5] proposed a novel approach that used a remote sensing top view image and three ground view images for joint land cover classification. Hong *et al.* [6] proposed a shared and specific feature learning model for land cover classification with multimodal remote sensing images, including hyperspectral images, SAR images, and DSM data.

Specifically, for joint land cover classification using optical and SAR images, studies started in the early 1990s, but the progress was slow. Zhang *et al.* [7] summarized the difficulties of joint classification into three problems. First, it is difficult to align optical and SAR images. As we have mentioned, optical and SAR images show different object features by individual imaging mechanisms, which enhances the difficulty of automatically selecting ground control points for registration, while manual selection is extremely inefficient. Second, there is no conclusion regarding which fusion level is the best. The existing approaches can be divided into three categories according to the content of the fusion. Pixel level uses the original images for fusion [8]. Feature level uses the classification features designed manually or extracted by the model for fusion [9].

Manuscript received August 31, 2021; revised December 16, 2021; accepted January 12, 2022. Date of publication January 21, 2022; date of current version February 9, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61901439 and in part by Key Research Program of Frontier Sciences, CAS, under Grant ZDBS-LY-JSC036. *(Corresponding author: Wenchao Kang.)*

Wenchao Kang, Yuming Xiang, and Hongjian You are with the Key Laboratory of Technology in Geo-Spatial Information Processing and Application System and the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Electronic, Electrical, and Communication Engineering, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: xshzhdm@163.com; z199208081010@163.com; hjyou@mail.ie.ac.cn).

Feng Wang is with the Key Laboratory of Technology in Geo-Spatial Information Processing and Application System and Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100190, China (e-mail: wfeng_gucas@126.com).

Digital Object Identifier 10.1109/JSTARS.2022.3144587

Decision level uses the classification results from each image for fusion [10]. Finally, which fusion strategy can fully exploit the complementarity of optical and SAR images remains unknown. Tupin and Florence [11] point out that commonly used optical image fusion methods may not be suitable for the fusion of optical and SAR images.

The heterogenous alignment of optical and SAR images is an important research area. Recent studies have included both traditional methods based on phase congruency [12] and alignment models based on deep learning [13]. As the matching problem is irrelevant to classification methods, the research usually focuses on the last two. To find the best fusion level, Zhang *et al.* [14] compared four fusion methods [maximum likelihood method (ML), artificial neural network (ANN), support vector machine (SVM), and random forest (RF)] in three fusion levels for a total of 12 different joint classification methods. Hong *et al.* [15] also explored various fusion levels with a patch-based convolutional neural network (CNN) for multiclass land cover classification. For the fusion method, Moumni *et al.* [16] and Hu *et al.* [17] compared traditional machine learning methods, including ANN, SVM, and ML. Lestari *et al.* [18] and Gbodjo *et al.* [19] explored the effects of CNNs for joint classification. Taking it a step further, Li *et al.* successively designed several fusion modules used in CNNs for joint classification, such as SACSM [20], MCAM, and GHFM [21]. Adrian *et al.* [22] utilized a 3-D U-Net for multitemporal crop type mapping.

The neural networks used for semantic segmentation have undergone a development process from ANN, CNN to FCN. The end-to-end structure of the FCN and its ability to handle arbitrary input sizes make it suitable for image segmentation tasks and greatly improve the efficiency of segmentation. FCNs have been widely used for heterogeneous remote sensing images joint land cover classification but do not include optical and SAR images. However, there have been many FCN-based land cover classification studies for single-source remote sensing images, either optical [23]–[25] or SAR [26], [27]. This is because it is barely usable optical and SAR image datasets to train an FCN. In addition, the graph convolution network has also been introduced for land cover classification [28].

Currently, FCN models are rarely applied to optical and SAR images joint land cover classification. In this article, we explore the effects of different fusion levels and fusion strategies on the classification results for the FCN model. The key to improving the classification accuracy is to make full use of the complementary characteristics in both images. Therefore, we design a novel FCN model for joint classification. To explore the effect of different fusion levels on the classification results, we carried out a modular design on the model. We divide the model into three parts: backbone, neck, and head, and compare the classification effects of four different fusion positions. To make better use of complementary features, we designed four different gate fusion modules and compared them with three basic fusion methods and one gate fusion module in other literature. In addition, considering the small amount of data available, we use a lightweight encoder and design a lightweight multiscale feature fusion module as the neck. To evaluate the effects, we finally implement exhaustive experiments on two datasets, one of which is our homemade dataset for multiclass land cover classification. Experimental results show that the cross-gate (CRG) fusion module has stronger generalizability than the other six modules. The contributions of this article are as follows.

1) To make better use of the FCN model in optical and SAR images joint land cover classification, we make a comprehensive discussion on the effects of different fusion positions and fusion strategies on the classification results.
2) We propose a modular network model that can flexibly adjust the fusion position and design a novel lightweight multiscale feature extraction module. In addition, we design four gate fusion modules and the CRG fusion module is the best performer.
3) We build a multiclass land cover classification dataset with Gaofen series satellite remote sensing images.

The rest of this article is organized as follows. Section II gives a detailed introduction of our model. The study data are described in Section III. The experimental results are illustrated and analyzed in Section IV. Finally, Section V concludes this article.

II. Method

We expect joint classification using optical and SAR images to achieve higher accuracy than the classification models using single-sensor images. For this purpose, we first design a modular model, which consists of three modules: encoder, multiscale module, and decoder, and allow the flexible replacement of arbitrary modules and adjustment of fusion approaches. Then, we investigate where and how to fuse optical and SAR features, and propose a CRG method with a bidirectional information flow to better utilize their complementary information.

A. Backbone

Due to the limited samples of the existing multisensor land cover classification datasets, we first design a self-attention multiscale network (SMNet) for small-sample land cover classification datasets, as the base model. The basic structure of the SMNet is shown in Fig. 1.

Based on the FCN [29], we design a lightweight model called a self-attention multiscale network (SMNet) in this article, as the base model. The complete model structure consists of three parts and is illustrated in Fig. 1.

Our target is to design a model suitable for land cover classification with a small dataset. In this case, a complex encoder such as VGG16 in the FCN is more likely to overfit with redundant features. Thus, we use a lightweight backbone MobileNetV2 [30] to replace VGG16. Besides, MobileNetV2 is deeper than VGG16 and can utilize high-level semantic features to increase the classification accuracy. Reducing the output stride (OS) is a commonly used method in semantic segmentation, as it is a location-sensitive task. However, due to speckle noise, a small output stride does not always work for SAR images. Therefore, we have 8x, 16x, and 32x alternative output strides and use 1, 2, and 3 transpose convolutions in the decoder, respectively. Short connections are also used for the last two

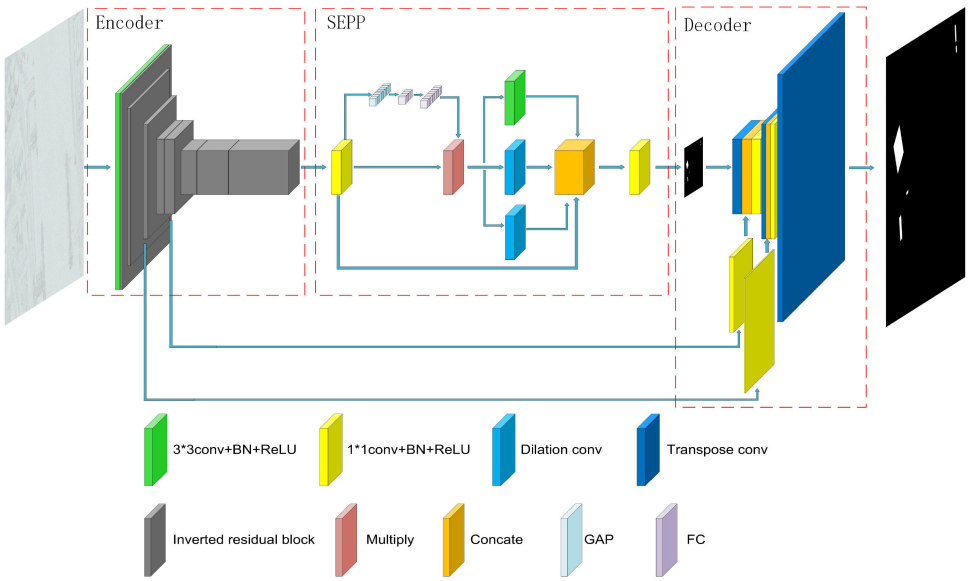

Fig. 1. Structure of SMNet.

situations. In addition, to offset the decrease in the receptive field, we replace the standard convolution in the last one or two blocks in the encoder, as the other research does. For the 16x output stride, the last block use rate=2. For the 8x output stride, the last two blocks use rate=2 and rate=4.

Between the encoder and decoder, Long *et al.* [29] added three extra convolutions in the FCN. The FCN is trained with 224*224 pixel images in the original article, and the output feature map of the encoder is 7*7 pixels after 32x downsampling. A 7*7 convolution following the encoder can extract global features, in which the high-level semantic information can help the model perform a more accurate classification. Subsequently, two 1*1 convolutions are used to integrate the global features and give the low-resolution classification result. Although these three convolutions further improve the classification results, they also produce a super large number of parameters, which are far more than the sum of encoder and decoder, due to the vast channels (4096). It is inefficient and inconsistent with our expectation of a lightweight model. If the training sample is over 224*224 pixels, the 7*7 convolution will lose its value for global feature extraction. Therefore, we design a novel squeeze-and-excitation pyramid pooling (SEPP) module to replace them in our model.

Considering that the problem of FCN uses too many channels, we first use a 1*1 convolution to condense the channels of the encoder output and extract the critical features. This is based on the assumption that encoding the manifolds of interest in a neural network only needs a low-dimensional subspace, proposed by [30] when designing MobileNetV2. Executing the subsequent operations on the subspace can effectively reduce the parameters and improve efficiency. To extract global features like the 7*7 convolution but efficiently, we use a squeeze-and-excitation (SE) [31] module to introduce the global information into the subspace. Besides, considering that different objects have various scales, we use three dilated convolutions to extract multiscale features, similar to the ASPP [32]. As the subspace already contains global information, we decide to use small rates 1, 3, and 6 for three dilated convolutions to extract more powerful local features. The dilated convolutions also use depthwise separable convolution to reduce parameters. To utilize more scale features at the same time, we stack the dilated convolution layers with the first 1*1 convolution layer and infer the classification result with a 1*1 convolution like FCN, as shown in Fig. 1. The reason for using the first 1*1 convolutional layer instead of the output of the SE module is that we want to retain the original features from the encoder and have a residual learning-like effect.

B. Fusion Position

The overall structure of SMNet can be divided into three modules: encoder, multiscale module, and decoder. When the optical and SAR images are fed into the network simultaneously, we can obtain four fusion positions: directly stack the raw images (input fusion), before the multiscale module (early fusion), after the multiscale module (late fusion), and after the decoder (output fusion).

1) *Input fusion:* The input fusion stacks two raw images, as shown in Fig. 2(a). Input fusion can retain the original information from two images. We can directly use the land cover classification model with the best performance on the single-source image.
2) *Early fusion:* Early fusion fuses the shallow layers of the model and shares the other modules. Of course, in a broad sense, input fusion can also be classified as early fusion. In this article, early fusion refers to the fusion of encoder output features, as shown in Fig. 2(b). Early fusion is feature-level fusion, in which the model uses the features of both images simultaneously for one decision.

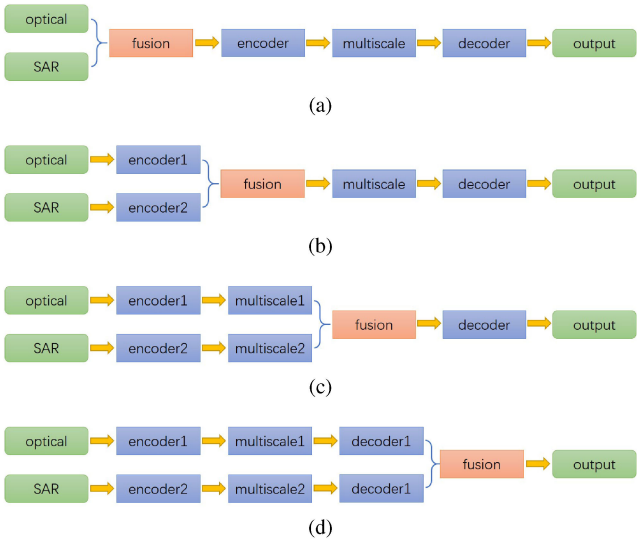

Fig. 2. Different fusion positions. (a) Input fusion. (b) Early fusion (c) Late fusion. (d) Output fusion.

It is suitable for a single land cover object that requires features from both images for classification.

3) *Late fusion:* Late fusion fuses deep layers of the model and shares the other modules. According to the type of deep layer output, late fusion can be either feature-level fusion or decision-level fusion. In this article, late fusion refers to the fusion of multiscale module output, as shown in Fig. 2(c). It is a decision-level fusion in which the model picks the exact classification from each of the two decision outcomes.
4) *Output fusion:* The output fusion fuses the classification results from two heterogeneous source images, as shown in Fig. 2(d). Output fusion is learnable postprocessing, which belongs to decision fusion. Similarly, in a broad sense, it can also be divided into late fusion.

The modules shared by the two branches are reduced sequentially in the aforementioned four fusion positions. Although we use SMNet as the backbone in this article, the proposed model is still flexible since other modules can be customized for heterogeneous images specifically.

C. Fusion Method

For input fusion, we directly stack the inputs to preserve the original information in two images. For the last three fusion positions, we hope to explore fusion methods that can improve the classification accuracy by fully utilizing the complementary information in two images. There are various fusion methods available, among which the most intuitive methods are add and stack. Add is to calculate the sum of two image features, as shown in Fig. 3(a). The stack can be divided into two types. The first is to pass directly to the subsequent network without postprocessing, as shown in Fig. 3(b). The second is to use 1×1 convolution for feature fusion after stacking, as shown in Fig. 3(c).

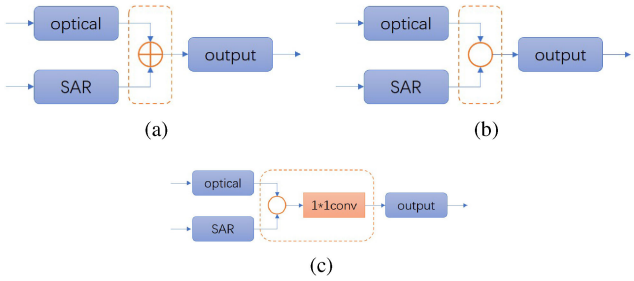

Fig. 3. Different kinds of basic fusion methods. (a) Add. (b) Stack1. (c) Stack2.

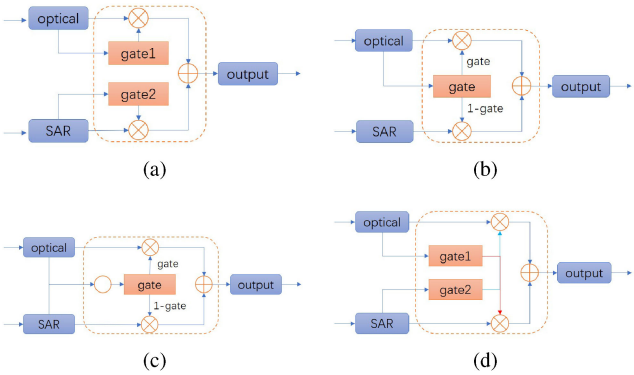

Fig. 4. Different kinds of gate fusion methods. (a) Independent gates (IGs). (b) Complementary gate1. (c) Complementary gate2. (d) Cross gates (CRGs).

The aforementioned three fusion methods merely utilize the features from two images simply and neglect to tap the complementary information in the two images. To solve this problem, we need to filter the redundant features. In the existing literature, the rules for feature screening are artificially set. It is neither possible to filter the best features nor to embed them in the FCN model. Therefore, the ideal outcome is that the network can learn how to filter on its own. For this purpose, we proposed four learnable gate modules, as shown in Fig. 4. All gate functions shown in Fig. 4 are SE modules.

First, the intuitive idea is to let the two images provide only the most useful features of each. A simple implementation is to learn a gate function for each image. As shown in Fig. 4(a), the two gates are independent of each other. We refer to this module as the independent gate (IG). The IG has an obvious problem, as there is no information exchange between the two gates. It cannot assure that the approved features are complementary.

To exploit the complementarity of the two images, we propose the second type of gate module, complementary gate1 (CG1), as shown in Fig. 4(b). CG1 learns only one gate from the main image. We compare the classification accuracy when using each single source image and specify the image with better results as the main image. In this article, it is the optical image. We use the gate and $1-$ gate to select the complementary features from two images. Although CG1 has information exchange between the two images, the unidirectional information flow increases the uncertainty of feature selection. Because the optical image can

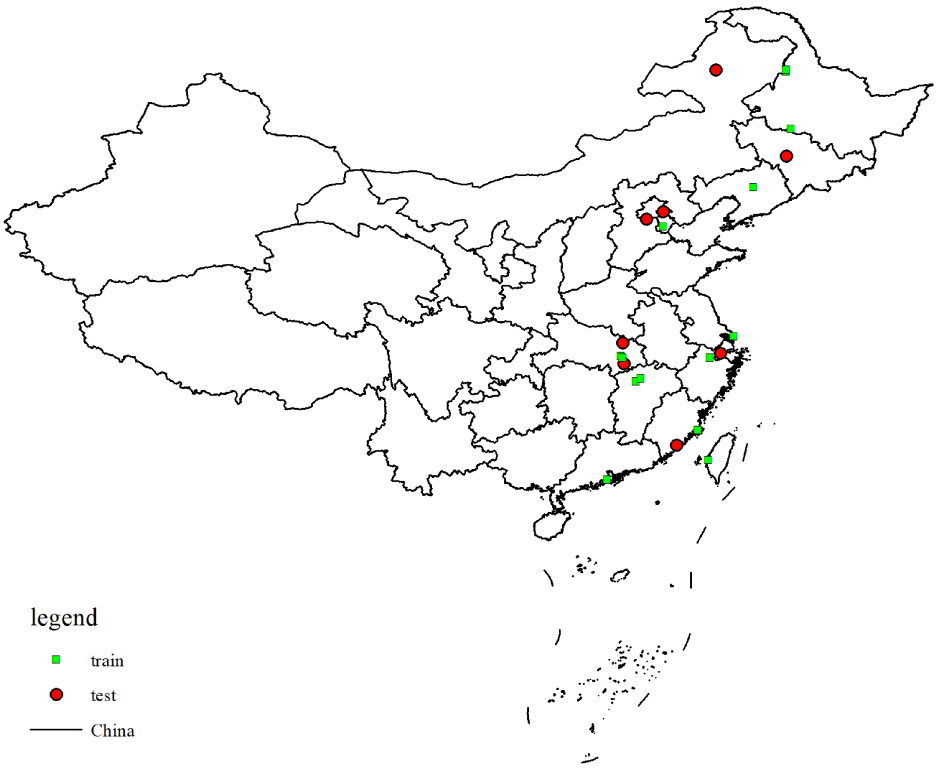

Fig. 5. Data distribution of the GMC dataset.

only guess which useful features should be retained and what complementary features the SAR image can provide.

To eliminate uncertainty, we need to construct a bidirectional information flow. If we drop the main image and instead learn the gate function from both images simultaneously, the gate can transmit the information bidirectionally. Based on CG1, we design complementary gate2 (CG2), as shown in Fig. 4(c). In CG2, we need to stack the optical image and SAR image first. However, there is still a problem for CG2 in which feature selection is either one or the other. When a common feature is of great help for classification, a better choice is to enhance it on both sides. In addition, this is something that CG2 cannot do.

To solve the aforementioned problem, a gate module must have two gates. Undoubtedly, bidirectional information flow is also necessary. Therefore, we construct the CRG, as shown in Fig. 4(d). A gate function is learned from each image separately. Unlike the IG, the gate function in the CRG learned in each image is used to control the other image, i.e., cross control. In this way, the CRG simultaneously obtains the advantages of both the IG and the CG.

III. Datasets

A. Gaofen Multiclass Dataset

To the best of our knowledge, for land cover classification, there is no multiclass object dataset using optical and SAR images. To conduct research, we construct a small dataset. Since there are many ready-made labels for optical remote sensing image datasets, we need to find the appropriate SAR images and match them with the optical images. To facilitate the retrieval of matching data, we finally choose the GID dataset [33] as the label source.

To reduce the matching difficulty and ensure the label accuracy as much as possible, we decided to search the SAR images with resolutions similar to those of the optical images. Because the GID dataset uses the Gaofen 2 multispectral images with a 3.24-m resolution, we chose Gaofen 3 ultrafine stripe (UFS) mode images with a 3-m resolution. There are 150 images in the GID dataset. After searching and filtering, we finally obtained 23 pairs of optical and SAR images with large overlapping areas. Among them, 15 are used for training and 8 for testing. We called it the Gaofen multiclass (GMC) dataset. The training set and test set use images from different cities without any overlapping areas, as shown in Fig. 5.

The GF2 optical images in the GID dataset do not have geocoding information. Therefore, to improve the matching efficiency and accuracy, we add a manual prematch process, and the complete matching process contains three steps. First, crop a prematched area of the same size as the optical image from the SAR image by manually selecting 4∼5 match points. Second, slice the prematched images into 1536*1536 pixel patches with a 768-pixel stride. Finally, we use the registration algorithm from [12] to match the slice pairs exactly and keep only the center 768*768 pixels. For ease of use, we crop the patch into 256×256 pixels without overlap. We obtain 7137 training samples and 3591 test samples. It is important to note that although the GID dataset provides six types of labels, there is almost no grass in our images. Thus, we decide to keep only the

TABLE I
MULTICLASS CLASSIFICATION RESULTS ON GMC DATASET WITH DIFFERENT MODELS (%)

model	encoder	OA	IoU					FWIoU
			Other	Built_up	Farmland	Forest	Water	
HR-SAR-Net		60.96	43.96	60.32	35.70	24.01	63.97	45.59
MPResNet	ResNet34	64.32	47.68	56.39	43.50	27.28	67.38	47.60
FCN-SAR	Vgg16	64.31	49.33	59.47	38.15	24.75	67.54	47.19
FCN8	Vgg16	64.82	47.88	59.25	40.88	30.99	**70.61**	48.02
PSPNet	MobileNetv2	65.88	45.92	61.76	**47.77**	34.59	70.05	49.52
DeepLabv3+	MobileNetv2	66.17	49.43	61.20	45.97	31.03	61.36	49.51
SMNet	MobileNetv2	**67.63**	**50.64**	**62.26**	46.44	**35.75**	69.34	**51.37**

Best in bold.

TABLE II
BUILDING EXTRACTION RESULTS ON SN6 DATASET WITH DIFFERENT MODELS (%)

model	encoder	OA	IoU
HR-SAR-Net		89.41	39.39
MPResNet	ResNet34	96.14	68.35
FCN-SAR	Vgg16	95.35	62.75
FCN8	Vgg16	95.23	62.71
PSPNet	MobileNetv2	96.07	68.04
DeepLabv3+	MobileNetv2	95.80	66.75
SMNet	MobileNetv2	**96.23**	**68.98**

Best in bold.

remaining five objects, i.e., Built_up, Farmland, Forest, Water, and Other.

B. SpaceNet6 Dataset

In 2020, SpaceNet released a multisensor all-weather mapping dataset SN6 for their 6th task [34]. SN6 is a building extraction dataset consisting of 3401 pair images with a size of 900×900 and 0.5-m resolution. Different from our GMC dataset, SN6 covers only Rotterdam, the Netherlands, an extent of approximately 120 km^2. However, heterogeneous geographies, such as high-density urban environments and rural farming areas, will also make classification difficult. Although the optical images in SN6 provide RGBNIR four bands, we only use RGB three bands in this article. Since the original data contain many invalid parts, i.e., regions with all zero values, we first used cropping to remove them and keep as many valid areas as possible. Then, we randomly divided images into a training set, validation set, and test set in an approximate ratio of 2:1:1. Samples with the same name and different serial numbers were divided into the same subset. Thus far, we have obtained 1711 training samples, 862 validation samples, and 828 test samples. To facilitate training, we sliced the training and validation sets without overlap. Considering that the data are relatively sufficient and the images with less than 1% of the building pixels are not very helpful for model training, we abandoned these slices to accelerate the training process. Finally, we obtained 11 485 training samples and 5 796 validation samples.

IV. EXPERIMENTS

A. Experimental Setting

Our model is built with Pytorch 1.6 and employs the pretrained MobileNetV2 on ImageNet. The optimizer is stochastic gradient descent (SGD) with a momentum of 0.9. The learning rate schedule is the poly policy.

$$\mathrm{lr} = (\mathrm{lr}_i - \mathrm{lr}_e) * \left(1 - \frac{\mathrm{itr}}{\max_{\mathrm{itr}}}\right)^{\mathrm{power}} + \mathrm{lr}_e \tag{1}$$

where the initial learning rate is 0.05 for GMC and 0.4 for SN6, the end learning rate is 0.001, and the power is 0.9. SMNet is trained on GMC for 40 epochs or on SN6 for 100 epochs. We train on two GPU with a batch size of 64. The loss function is cross-entropy loss.

We adopt two metrics for the evaluation of classification results. They are overall accuracy (OA) and intersection over union (IoU). For GMC, we introduce an additional metric frequency weighted IoU (FWIoU). Let p_{ij} be the number of pixels in class i predicted to belong to class j, where there are k different classes. We compute:

$$\mathrm{OA} = \frac{\sum_{i=1}^{k} p_{ii}}{\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}} \tag{2}$$

$$\mathrm{IoU} = \frac{p_{ii}}{\sum_{j=1}^{k} p_{ij} + \sum_{j=1}^{k} p_{ji} - p_{ii}} \tag{3}$$

TABLE III
CLASSIFICATION RESULTS OF USING SINGLE SOURCE IMAGES FOR GMC DATASET (%)

data \ OS	8x		32x	
	OA	FWIoU	OA	FWIoU
optical	72.16	56.89	72.06	56.60
SAR	66.69	50.14	67.63	51.37
Δ	5.47	6.75	4.43	5.23

TABLE IV
CLASSIFICATION RESULTS OF USING SINGLE-SOURCE IMAGES FOR SN6 DATASET (%)

data \ OS	8x		32x	
	OA	IoU	OA	IoU
optical	98.31	85.13	98.06	83.13
SAR	96.23	68.98	95.65	64.84
Δ	2.09	16.15	2.42	18.29

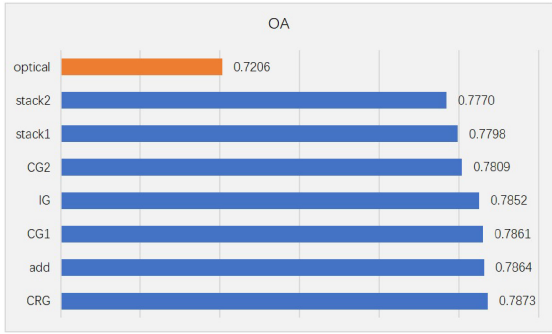

Fig. 6. Comparison of OA on the GMC dataset using different fusion modules.

TABLE V
EXPERIMENTS RESULTS OF GMC DATADET FOR USING DIFFERENT FUSION POSITIONS (%)

position	OA	IoU					FWIoU
		other	built-up	farmland	forest	water	
input	72.35	53.90	66.34	52.82	45.07	84.82	57.05
early	78.12	61.14	68.71	65.59	47.70	89.03	64.36
late	78.64	61.64	69.49	65.18	52.39	89.95	65.08
output	77.91	60.49	68.03	64.88	49.96	89.14	64.01

TABLE VI
EXPERIMENTS RESULTS OF SN6 DATADET FOR USING DIFFERENT FUSION SITES (%)

position	OA	IoU
input	98.00	82.60
early	98.09	83.41
late	98.07	83.13
output	98.00	82.53

TABLE VII
EXPERIMENTS RESULTS OF GMC DATADET FOR USING DIFFERENT FUSION MODULES (%)

module	OA	FWIoU
stack1	77.98	64.31
stack2	77.70	63.80
add	78.64	65.08
IG	78.52	64.95
CG1	78.61	65.05
CG2	78.09	64.40
CRG	78.73	65.20

TABLE VIII
EXPERIMENTS RESULTS OF SN6 DATASET FOR USING DIFFERENT FUSION MODULES (%)

module	OA	IoU
stack1	98.10	83.34
stack2	98.10	83.42
add	98.09	83.41
IG	98.10	83.39
CG1	98.06	82.99
CG2	98.05	83.04
CRG	98.11	83.47

$$\text{FWIoU}=\frac{1}{\sum_{i=1}^{k}\sum_{j=1}^{k}p_{ij}}\sum_{i=1}^{k}\frac{p_{ii}}{\sum_{j=1}^{k}p_{ij}+\sum_{j=1}^{k}p_{ji}-p_{ii}}. \tag{4}$$

B. Base Model

As SMNet is used as the base model for joint land cover classification, we first assess the performance of SMNet with only SAR images. We compare it with six models. HR-SAR-Net [35], FCN-SAR [26], and MP-ResNet [27] three models are especially designed for land cover classification of SAR images. The three models FCN8 [29], DeepLabv3+ [32], and PSPNet [36] are classical semantic segmentation models for natural images.

The multiclass land cover classification results of each model on the GMC dataset are shown in Table I. HR-SAR-Net is the worst performer, and the gap with other models is obvious. MP-ResNet outperforms FCN-SAR by a little bit but performs much worse than the three classical models. Although PSPNet does not perform as well as DeepLabv3+ in OA, it performs better in FWIoU. There is no doubt that our model achieves the best performance, improving the FWIoU metric by 1.85% compared to the suboptimal performing PSPNet model.

The building extraction results on the SN6 dataset are shown in Table II. HR-SAR-Net gives the worst performance again with a huge gap compared to the other models. This is because HR-SAR-Net is too simple to extract enough powerful features as SN6 is relatively large. For this time, MP-ResNet has better performance than PSPNet and DeepLabv3+. Ding *et al.* [27] designed MP-ResNet for land cover classification with full-polarization SAR images, to which the SN6 data exactly belong. For this time, PSPNet outperforms DeepLabv3+, which is consistent with the fact that the PPM module performs better than the ASPP module in our model. There is no doubt that our model SMNet performs best overall. Compared with MP-ResNet, SMNet improves the IoU metric by 0.63%.

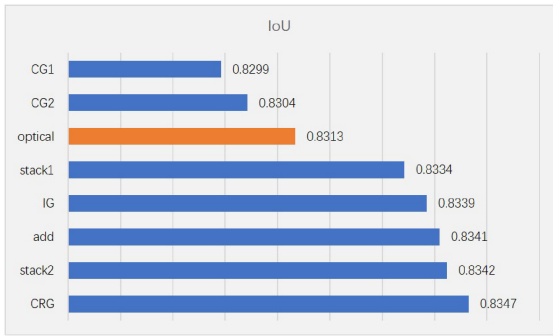

Fig. 7. Comparison of accuracy metrics on SN6 dataset for using different fusion modules.

TABLE IX
EXPERIMENTAL RESULTS ON GMC DATASET AND SN6 DATASET FOR USING DIFFERENT OUTPUT STRIDE (%)

OS	GMC		SN6	
	OA	FWIoU	OA	IoU
8x	79.05	65.66	98.33	85.26
16x	77.75	63.84	98.13	83.59
32x	78.73	65.20	98.11	83.47

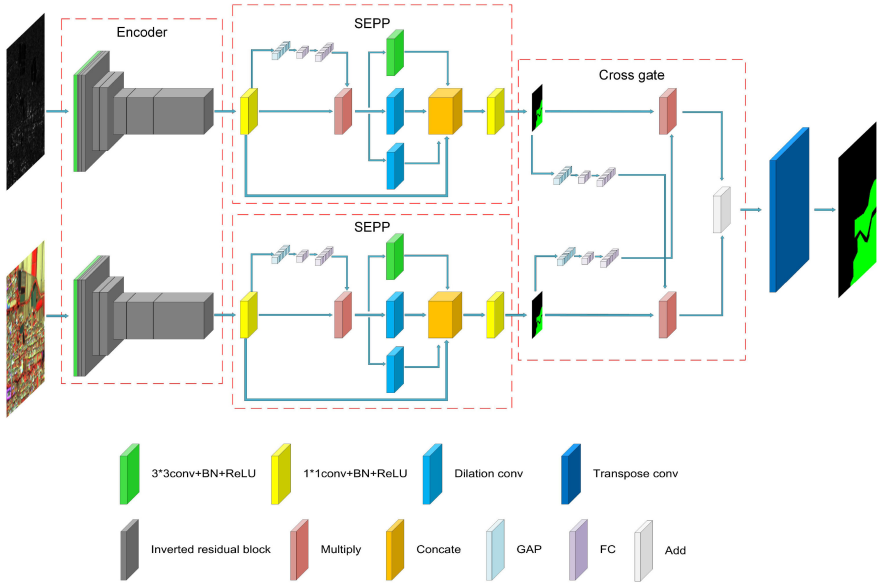

Fig. 8. Structure of CFNet with late fusion.

TABLE X
EXPERIMENTAL RESULTS ON GMC DATASET AND SN6 DATASET FOR ADDING SHORT CONNECTION (SC) (%)

SC	GMC		SN6	
	OA	FWIoU	OA	IoU
0	79.05	65.66	98.33	85.26
1	78.43	64.82	98.37	85.61
2	78.50	64.92	98.60	87.47

C. Baseline

To verify the effectiveness of the joint classification, we need to set a baseline. Thus, we first evaluate the classification results of using single-source images separately. We use SMNet as the base model and test the classification accuracies with 8x and 32x two output strides. Tables III and IV give the metrics on the GMC and SN6 datasets, respectively. For the convenience of analysis, we also give the difference Δ between optical and SAR image indicators.

The classification metrics of both optical and SAR images in the GMC dataset are lower, but the Δ is also smaller. In contrast, the classification metrics of both optical and SAR images in the SN6 dataset are higher, and the Δ is also very large. The Δ of IoU is even near 20%. The improvement of joint classification on the SN6 dataset will be rather limited predictably. As the optical image outperformed the SAR image on both datasets, we use the former as the baseline for the subsequent experiments. Besides, while reducing the OS, the optical image classification metrics have more or less improvement for both datasets.

D. Fusion Position

We first perform comparison experiments of fusion positions to determine the overall structure of the network. Except for the input fusion, the fusion methods uniformly use the add to simplify the comparison. To save time, we set the OS to 32x.

The classification results of the GMC dataset are shown in Table V. Compared with the baseline, the different joint classification models with all four fusion positions have improved classification metrics. Excluding input fusion, all fusion methods showed significant improvement. Among them, late fusion gave the best performance, with a 6.58% improvement in OA and an 8.48% improvement in FWIoU. The early fusion had a very close performance. Although the input fusion improved the overall accuracy, the IoU of built-up, farmland, and water instead decreased.

The classification results of the SN6 dataset are shown in Table VI. Compared with the baseline, even the best-performing early fusion can only achieve a negligible improvement with a 0.03% improvement in OA and a 0.28% improvement in IoU. Using input fusion and output fusion even made the classification results worse. This result confirms our previous conjecture that the joint classification will not perform well on the SN6 dataset.

Taken together, although input fusion can reuse the model designed for a single image, it cannot exploit the complementarity from two images. Thus, input fusion is helpless for joint classification. Output fusion is more like postprocessing, which consumes twice the resources, but the performance is not stable. In contrast, early fusion and late fusion achieve relatively stable improvements even on SN6, a difficult dataset for joint classification. Based on the results in Tables V and VI, in subsequent experiments, the GMC dataset uses late fusion, while the SN6 dataset uses early fusion.

E. Fusion Module

After determining the fusion positions, we next compare the classification effects of the seven fusion modules. The experimental results for the GMC dataset are shown in Table VII.

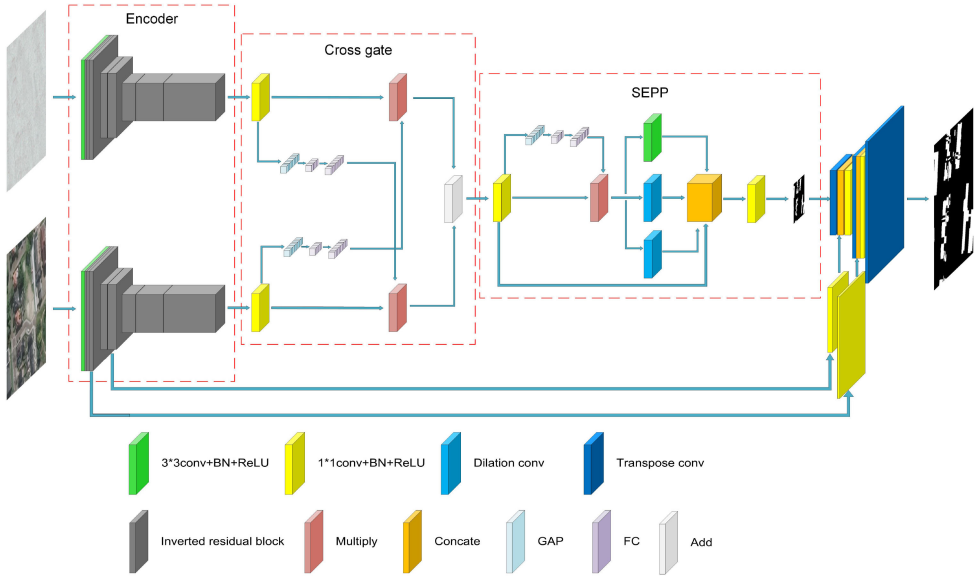

Fig. 9. Structure of CFNet with early fusion.

TABLE XI
EXPERIMENT RESULTS OF TWO MODELS ON GMC DATASET (%)

model	OA	IoU					FWIoU	params	FLOPs	FPS
		other	built-up	farmland	forest	water				
CFNet	79.05	61.25	70.95	66.24	55.98	89.47	65.66	4.77M	5.11G	41
VFuseNet	77.37	54.34	63.31	51.81	47.34	86.31	56.89	41.36M	54.99G	48

TABLE XII
EXPERIMENT RESULTS OF TWO MODELS ON SN6 DATASET (%)

model	OA	IoU	params	FLOPs	FPS
CFNet	98.60	87.47	5.24M	5.19G	6
VFuseNet	97.22	76.60	41.36M	54.99G	2

For a clear comparison, the OAs of different fusion modules and the baseline are shown in Fig. 6. We can see that each fusion module has a substantial improvement compared to the baseline. The four gate modules perform as expected, except for CG2. The classification accuracy gradually improves from the IG using no information exchange to the CG1 using unidirectional information flow to the CRG using bidirectional information flow. In contrast, CG2, which also uses bidirectional information flow, does not work as expected. Given that the two stack modules have the worst performance among all modules, perhaps the stack is not conducive for the GMC dataset. Besides, add is simple but surprisingly achieves the second-best classification results, after our especially designed CRG.

The experimental results of the SN6 dataset are shown in Table VIII. Similarly, a comparison of IoU for different fusion modules and baseline are shown in Fig. 7. Unlike the GMC dataset, not all fusion modules improve the classification results on the SN6 dataset. However, what remains the same is that the CRG is still the best performing module, which is in line with our expectations. For the four gate modules, two CG modules containing only one gate perform worse than the IG module and the CRG module that uses two gates. As we have mentioned before, this is because the CG cannot enhance a common essential feature on both sides at the same time.

Although the performances of the four gate modules are not completely consistent with our expectations, the most important CRG module gives satisfactory results. And, we will use it in the following experiments.

F. Output Stride

In the previous experiments, we used encoders with 32x downsampling to improve the experimental efficiency. We know that decreasing the output stride can preserve more detailed information and may improve the classification accuracy. In the baseline experiments, we found that reducing the output stride did improve the classification accuracy, except for the SAR image in GMC. Thus, we experimentally verified the effect of the encoder output stride on joint classification accuracy. The experimental results are shown in Table IX.

As the output stride decreases, the metrics first decrease, and then, increase for the GMC dataset. This is because the SAR images obtain the highest accuracy with the 32x output stride. Although the decrease in SAR metrics is larger than the increase in optical metrics in baseline experiments, the 8x output stride has a slight improvement over the 32x output stride for joint classification. If efficiency is preferential, 32x output stride is

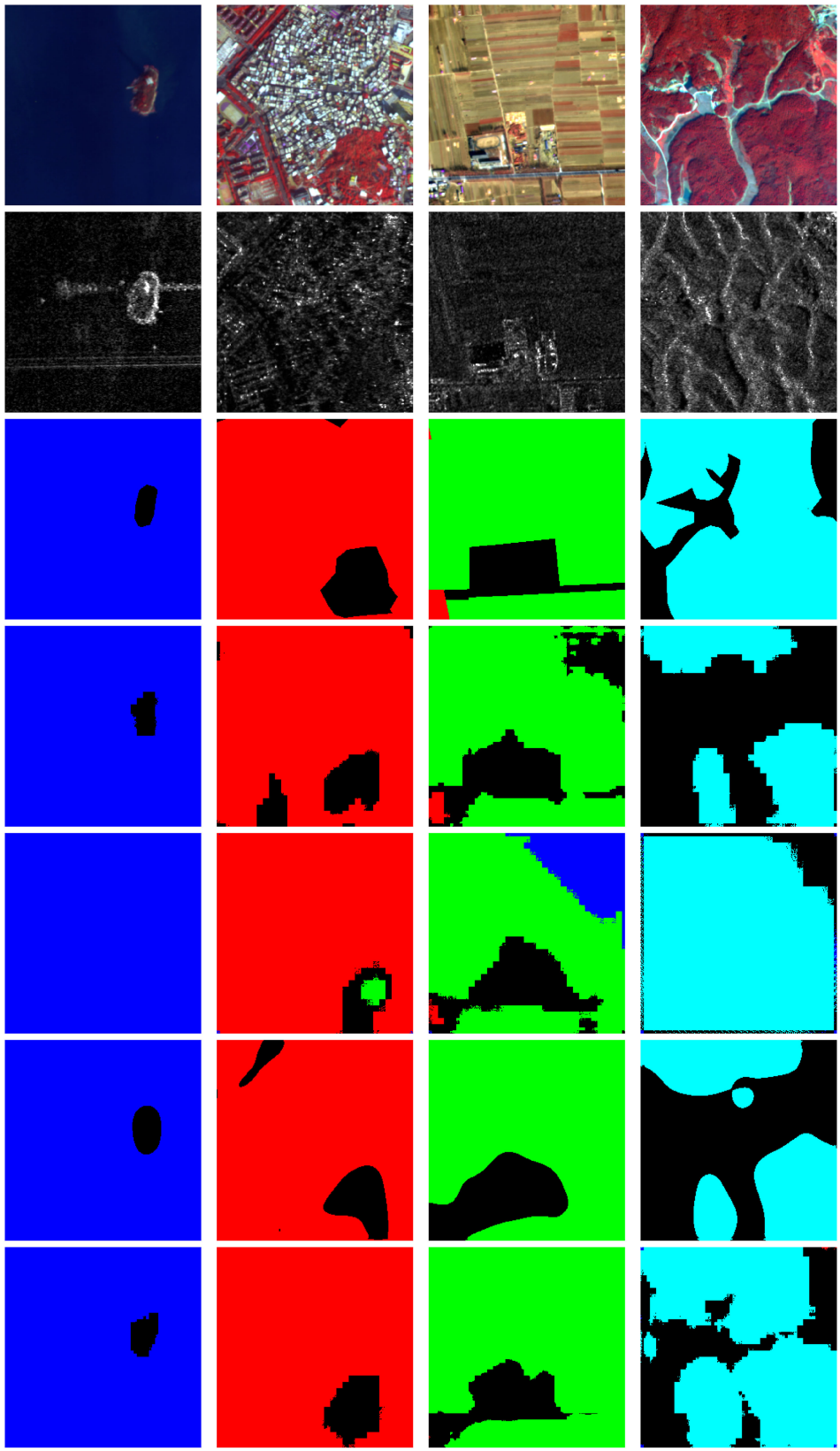

Fig. 10. Some visual results of the GMC dataset. (a) Water. (b) Built-up. (c) Farmland. (d) Forest. From top to bottom, they are original optical image, original SAR image, label, prediction of optical image, prediction of SAR image, VFuseNet, and CFNet.

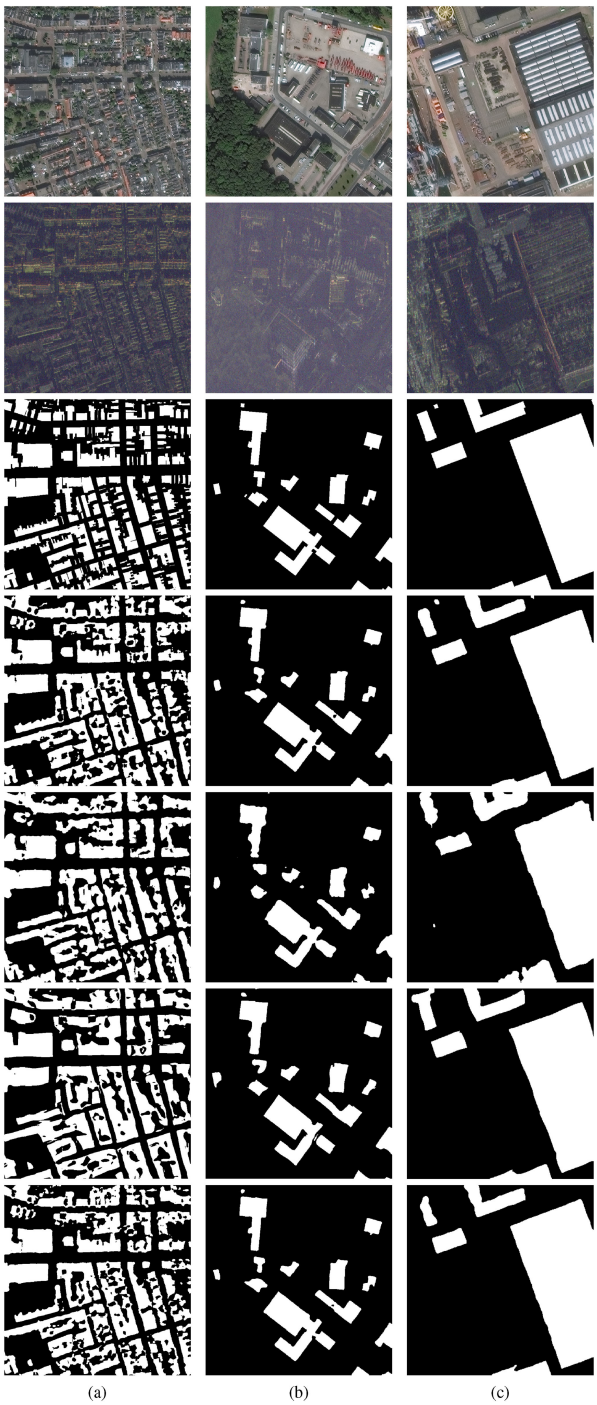

Fig. 11. Some visual results of the SN6 dataset. (a) Small buildings. (b) Medium buildings. (c) Large buildings. From top to bottom, they are original optical image, original SAR image, label, prediction of optical image, prediction of SAR image, VFuseNet, and CFNet.

a better choice. Here, we want to obtain higher classification accuracy, so an 8x output stride is preferable.

Unlike the GMC dataset, the classification metrics continuously improve as the output stride decreases for the SN6 dataset. The 8x output stride has an obvious improvement over the 32x output stride. However, compared to using only optical images, the advantage of joint classification is decreasing. We believe this is because the accuracy of the optical images is already high enough.

In summary, an 8x output stride is a better choice for both datasets to achieve a higher classification accuracy.

G. Decoder

The short connection between the encoder and the decoder is a commonly used method to transmit detailed information to help classification. The speckle noise in SAR images is an obstruction for land cover classification and may likewise be reintroduced into the decoder by short connection. Therefore, we only used a transposed convolution with 8x upsampling for 8x output stride in the SMNet and followed this structure for joint classification in previous experiments. However, for joint classification, we also have optical images without speckle noise. Therefore, we decided to verify whether we can obtain better performance by adding more short connections, as shown in Table X.

Unfortunately, adding the short connection cannot improve the classification results for the GMC dataset. We think this is because using an 8x output stride has already reserved enough detailed information for the five classification objects in the GMC dataset. Thus far, we have finally determined the best suitable structure for the GMC dataset, containing two MobileNetv2 with 8x output stride, two SEPP, one CRG fusion module, and one transposed convolution with 8x upsampling, as shown in Fig. 8. First, the two encoders and the SEPP module give the respective low-resolution classification results of optical and SAR images. Then, the CRG module is used to fuse the two to obtain the final low-resolution classification results. Finally, a transposed convolution with 8x upsampling is used to obtain the original resolution classification results.

For the SN6 dataset, the classification metrics have a sequential improvement while adding more short connections. As each building has a clear contour, more detailed information can help acquire more precise boundaries. Therefore, the best suitable structure for the SN6 dataset contains two MobileNetv2 with an 8x output stride, one CRG fusion module, one SEPP, and one decoder with three transposed convolutions and two short connections, as shown in Fig. 9. First, the classification features are extracted from the optical image and SAR image by the two encoders. Then, useful or complementary features are filtered from two sides using the CRG module and passed through the SEPP module to obtain the low-resolution classification results. Finally, three transposed convolutions with 2x upsampling are used to obtain the classification results at the original resolution.

Although the fusion positions are different, we refer to the aforementioned two models collectively as the cross fusion network CFNet.

H. Comparison With Other Model

Since there is no publicly available FCN model for the joint classification of optical and SAR images, we chose a model designed for optical images and DSM data as a reference. Audebert *et al.* [3] proposed two structures for fully convolutional models, VFuseNet and SegNet-RC belong to early fusion and output

fusion according to our division method. We choose VFuseNet, which performs better in the original article.

The metric results on the GMC dataset and the SN6 dataset are shown in Tables XI and XII. Our CFNet performs significantly better than VFuseNet on both datasets, and the latter even performs worse on the SN6 dataset than the optical baseline. In addition, our CFNet is a lightweight model designed for the joint classification of optical and SAR images with small samples. Taking the model used on the GMC dataset as an example, CFNet has only 4.77 M parameters, while VFuseNet has 41.36 M parameters. Our model is small but effective. However, the efficiency of our CFNet has no advantage. One reason is that our model is deeper. Besides, our model uses many separable convolutions that need to perform more memory reads and writes than standard convolutions.

The results of the GMC dataset are shown in Fig. 10. From top to bottom are the optical image, SAR image, label, optical image prediction result, SAR image prediction result, VFuseNet, and CFNet. The results improve significantly after using joint classification, both for VFuseNet and our CFNet. In comparison, the overall view of our CFNet is more accurate, while the edges of VFuseNet are smoother.

The results of SN6 are shown in Fig. 11. From top to bottom, they are optical image, SAR image, label, optical image prediction result, SAR image prediction result, VFuseNet, and CFNet. Optical image results are significantly better than the SAR image results, especially in terms of contour accuracy. However, it is difficult to perceive the improvement of the joint classification from the figure because the accuracy of the optical image is already quite high.

V. Conclusion

In this article, we investigated where and how to fuse optical and SAR images for joint classification. We designed a modular FCN CFNet to flexibly adjust the fusion position for different datasets. We compared four positions and found that early fusion and late fusion had stable performance. When the difference between the classification accuracy of two images is large, early fusion has better performance, otherwise, late fusion performs better. We designed a CRG fusion module that enables a bidirectional information flow and cross control through two gates. Therefore, it can better preserve important or complementary features.

Acknowledgment

The authors would like to thank the team of L. Zhang for opening the GID dataset and SpaceNet for opening the SN6 dataset.

References

[1] D. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[2] M. Volpi and D. Tuia, "Deep multi-task learning for a geographically regularized semantic segmentation of aerial images," *ISPRS J. Photogramm. Remote Sens.*, vol. 144, pp. 48–60, 2018.

[3] N. Audebert, B. Le Saux, and S. Lefèvre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks," *ISPRS J. Photogramm. Remote Sens.*, vol. 140, pp. 20–32, 2018.

[4] Y. Xu *et al.*, "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 6, pp. 1709–1724, Jun. 2019.

[5] S. Srivastava, J. E. Vargas-Munoz, and D. Tuia, "Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution," *Remote Sens. Environ.*, vol. 228, pp. 129–143, 2019.

[6] D. Hong, J. Hu, J. Yao, J. Chanussot, and X. X. Zhu, "Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model," *ISPRS J. Photogramm. Remote Sens.*, vol. 178, pp. 68–80, 2021.

[7] H. Zhang *et al.*, "A manifold learning approach to urban land cover classification with optical and radar data," *Landscape Urban Plan.*, vol. 172, pp. 11–24, 2018.

[8] L. Jiang, M. Liao, H. Lin, and L. Yang, "Synergistic use of optical and InSAR data for urban impervious surface mapping: A case study in Hong Kong," *Int. J. Remote Sens.*, vol. 30, no. 11, pp. 2781–2796, 2009.

[9] Y. Zhang, H. Zhang, and H. Lin, "Improving the impervious surface estimation with combined use of optical and SAR remote sensing images," *Remote Sens. Environ.*, vol. 141, pp. 155–167, 2014.

[10] H. Zhang, H. Lin, Y. Li, Y. Zhang, and C. Fang, "Mapping urban impervious surface with dual-polarimetric SAR data: An improved method," *Landscape Urban Plan.*, vol. 151, pp. 55–63, 2016.

[11] F. Tupin, "Fusion of optical and SAR images," in *Radar Remote Sensing of Urban Areas*. Berlin, Germany: Springer, 2010, pp. 133–159.

[12] Y. Xiang, R. Tao, F. Wang, H. You, and B. Han, "Automatic registration of optical and SAR images via improved phase congruency model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5847–5861, 2020.

[13] L. H. Hughes, D. Marcos, S. Lobry, D. Tuia, and M. Schmitt, "A deep learning framework for matching of SAR and optical imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 169, pp. 166–179, 2020.

[14] H. Zhang and R. Xu, "Exploring the optimal integration levels between SAR and optical data for better urban land cover mapping in the pearl river delta," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 64, pp. 87–95, 2018.

[15] D. Hong *et al.*, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.

[16] A. Moumni and A. Lahrouni, "Machine learning-based classification for crop-type mapping using the fusion of high-resolution satellite imagery in a semiarid area," *Scientifica*, vol. 2021, 2021, Art. no. 8810279.

[17] B. Hu *et al.*, "Improving urban land cover classification with combined use of sentinel-2 and sentinel-1 imagery," *ISPRS Int. J. Geo- Inf.*, vol. 10, no. 8, pp. 533–548, 2021.

[18] A. I. Lestari, M. Rizkinia, and D. Sudiana, "Evaluation of combining optical and SAR imagery for burned area mapping using machine learning," in *Proc. IEEE 11th Annu. Comput. Commun. Workshop Conf.*, 2021, pp. 00 52–0059.

[19] J. E. Gbodjo, O. Montet, D. Ienco, R. Gaetano, and S. Dupuy, "Multisensor land cover classification with sparsely annotated data based on convolutional neural networks and self-distillation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 11485–11499, 2021.

[20] X. Li, L. Lei, Y. Sun, M. Li, and G. Kuang, "Multimodal bilinear fusion network with second-order attention-based channel selection for land cover classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1011–1026, 2020.

[21] X. Li, L. Lei, Y. Sun, M. Li, and G. Kuang, "Collaborative attention-based heterogeneous gated fusion network for land cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 3829–3845, May 2021.

[22] J. Adrian, V. Sagan, and M. Maimaitijiang, "Sentinel SAR-optical fusion for crop type mapping using deep learning and Google Earth Engine," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 215–235, 2021.

[23] S. Dong, Y. Zhuang, Z. Yang, L. Pang, H. Chen, and T. Long, "Land cover classification from VHR optical remote sensing images by feature ensemble deep learning network," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1396–1400, Aug. 2020.

[24] Y. Liu, B. Fan, L. Wang, J. Bai, S. Xiang, and C. Pan, "Semantic labeling in very high resolution images via a self-cascaded convolutional neural network," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 78–95, 2018.
[25] A. Sellami and S. Tabbone, "Deep neural networks-based relevant latent representation learning for hyperspectral image classification," *Pattern Recognit.*, vol. 121, 2022, Art. no. 108224.
[26] W. Kang, Y. Xiang, F. Wang, L. Wan, and H. You, "Flood detection in Gaofen-3 SAR images via fully convolutional networks," *Sensors*, vol. 18, no. 9, pp. 2915–2936, 2018.
[27] L. Ding *et al.*, "Mp-ResNet: Multi-path residual network for the semantic segmentation of high-resolution PolSAR images," *IEEE Geosci. Remot. Sens. Lett.*, vol. 19, May 2021, Art. no. 4014205, doi: 10.1109/LGRS.2021.3079925.
[28] L. Gao, D. Hong, J. Yao, B. Zhang, P. Gamba, and J. Chanussot, "Spectral superresolution of multispectral imagery with joint sparse and low-rank learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2269–2280, Mar. 2021.
[29] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
[30] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.
[31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[32] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
[33] X.-Y. Tong *et al.*, "Learning transferable deep models for land-use classification with high-resolution remote sensing images," Jul. 2018, *arXiv:1807.05713*.
[34] J. Shermeyer *et al.*, "Spacenet 6: Multi-sensor all weather mapping dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 196–197.
[35] X. Wang, L. Cavigelli, M. Eggimann, M. Magno, and L. Benini, "HR-SAR-Net: A deep neural network for urban scene segmentation from high-resolution SAR data," in *Proc. IEEE Sensors Appl. Symp.*, 2020, pp. 1–6.
[36] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

Wenchao Kang received the B.S. degree in communication engineering from the Beijing Institute of Technology, Beijing, China, in 2016, and the Ph.D. degree in information and signal processing from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2021.

He is currently an Assistant Researcher with the Research Center for Brain-Inspired Intelligence, Institute of Automation, Chinese Academy of Science. His current research interests include remote sensing image land cover classification and object detection.

Yuming Xiang (Member, IEEE) received the B.S. degree in electronic engineering from the Tsinghua University, Beijing, China, in 2013, and the Ph.D. degree in information and signal processing from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2018.

He is currently an Assistant Researcher with the Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, Aerospace Information Research Institute, Chinese Academy of Science. His current research interests include remote sensing image registration and 3-D reconstruction.

Feng Wang received the B.S. degree in photoelectric information engineering from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2010, and the Ph.D. degree in signal and information processing from the University of Chinese Academy of Sciences, Beijing, in 2015.

Since 2015, he has been an Assistant Researcher with the Aerospace Information Research Institute, Chinese Academy of Sciences. His current research interests include multisource remote sensing image processing, image registration, and change detection.

Hongjian You received the B.S. degree in engineering from Wuhan University, Wuhan, China, in 1992, the M.S. degree in engineering from Tsinghua University, Beijing, China, in 1995, and the Ph.D. degree in remote sensing from the University of Chinese Academy of Sciences, Beijing, in 2001.

He is currently a Professor with the Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, Aerospace Information Research Institute, Chinese Academy of Science. His main research interests include remote sensing image processing and analysis and synthetic aperture radar image applications.