IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 26, 2020, accepted June 12, 2020, date of publication June 22, 2020, date of current version July 3, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3003914

Semantic Segmentation of Remote Sensing Images Using Transfer Learning and Deep Convolutional Neural Network With Dense Connection

BINGE CUI, XIN CHEN, AND YAN LU

College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China

Corresponding author: Yan Lu (luyan@sdust.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFC1405600, and in part by the National Natural Science Foundation of China (NSFC) under Grant 41406200 and Grant 41706105.

ABSTRACT Semantic segmentation is an important approach in remote sensing image analysis. However, when segmenting multiobject from remote sensing images with insufficient labeled data and imbalanced data classes, the performances of the current semantic segmentation models were often unsatisfactory. In this paper, we try to solve this problem with transfer learning and a novel deep convolutional neural network with dense connection. We designed a UNet-based deep convolutional neural network, which is called TL-DenseUNet, for the semantic segmentation of remote sensing images. The proposed TL-DenseUNet contains two subnetworks. Among them, the encoder subnetwork uses a transferring DenseNet pretrained on three-band ImageNet images to extract multilevel semantic features, and the decoder subnetwork adopts dense connection to fuse the multiscale information in each layer, which can strengthen the expressive capability of the features. We carried out comprehensive experiments on remote sensing image datasets with 11 classes of ground objects. The experimental results demonstrate that both transfer learning and dense connection are effective for the multiobject semantic segmentation of remote sensing images with insufficient labeled data and imbalanced data classes. Compared with several other state-of-the-art models, the kappa coefficient of TL-DenseUNet is improved by more than 0.0752. TL-DenseUNet achieves better performance and more accurate segmentation results than the state-of-the-art models.

INDEX TERMS Dense connection, transfer learning, remote sensing image, multiscale feature fusion, semantic segmentation, UNet.

I. INTRODUCTION

With the rapid development of remote sensing technology, a massive number of remote sensing images are becoming available every day [1]. Semantic segmentation, which aims at the pixel-level classification of images, has become an urgent need [2]. Semantic segmentation is one of the fundamental ways to analyze remote sensing images. This approach can easily and quickly obtain the land cover information of the area of interest, thereby providing data support for applications such as precision agriculture, desertification detection, traffic supervision, urban planning, and land resource management [3]–[9].

In recent years, deep convolutional neural networks have achieved great success in many fields, and have proven their excellent performance in many applications [10]. This trend has also attracted many researchers to apply deep convolutional neural networks to the field of remote sensing image semantic segmentation [11]–[13]. The fully convolutional neural network (FCN) [14] and its variants have exhibited excellent segmentation abilities. Sherrah [15] used an FCN-based network without any downsampling to semantically segment high-resolution aerial images. Their method used dilated convolution in DeepLab [16] which can maintain the full resolution of the images in each layer of the network

The associate editor coordinating the review of this manuscript and approving it for publication was Stefania Bonafoni.

and make better use of image features. Compared with the original FCN, this method has no downsampling layer, and the segmentation accuracy is higher. Bittner *et al.* [17] proposed the Fused-FCN4s model consisting of three parallel FCN4s networks. Three-band (R, G, B), PAN (panchromatic) and nDSM (normalized digital surface model) images were used as inputs to the parallel networks to extract features from high-resolution remote sensing images. Chen *et al.* [18] proposed a symmetrical FCN model, including the symmetrical normal shortcut FCN (SNFCN) and the symmetrical dense-shortcut FCN (SDFCN) with a shortcut connection. This structure outperformed the traditional methods, and has a symmetrical encoder and decoder, which solves the problem that the structure of the decoder is always simpler and shallower than that of the encoder.

Although the various FCN-based methods mentioned above have achieved remarkable performances in the field of remote sensing image segmentation, their recognition capabilities rely heavily on the large-scale dataset [19], since there are millions of parameters in the network that need to be trained. For remote sensing images with insufficient labeled data, previous studies have mainly focused on data augmentation [20] or designing relatively uncomplicated networks to avoid overfitting [21]. However, recent studies have indicated that the deeper the network is, the better the performance of the deep convolutional network [22]. Unfortunately, as the number of neural network layers increases, vanishing gradients problem may emerge. Thus, insufficient labeled data and vanishing gradients problem are the main obstacles to training deep convolutional neural networks for remote sensing image segmentation.

To address the first problem mentioned above, transfer learning, as a strategy of deep learning, provides an effective way to train a large network with limited data without overfitting. Yosinski *et al.* [23] experimentally quantified the generality versus specificity of neurons in each layer of a deep convolutional neural network and verified that transferring features even from distant tasks yields better performance than using random features. In addition, many recent studies [24]–[27] have demonstrated that deep convolutional networks pretrained on large natural image datasets such as ImageNet [28] can be transferred to other datasets with insufficient labeled data and perform better than other deep learning methods. Marmanis *et al.* [29] exploited a pretrained convolutional neural network based on ImageNet to extract the initial set of representations, and then transferred it to a supervised convolutional neural network classifier. Their best result over the UC Merced Land Use benchmark improved the overall accuracy (OA) from 83.1% to 92.4%, indicating that transferring representations from different fields may also be well suited for remote sensing image classification tasks. ImageNet is widely used as the source dataset in transfer learning cases due to its large amount of labeled data.

To solve the second problem above, He *et al.* [30] proposed ResNet with typical residual connection, which allows the gradient to flexibly propagate through the bypassing paths. Huang *et al.* [31] proposed DenseNet, which utilizes a dense connection method to cope with the vanishing gradients problem. Through dense connection, each convolutional layer receives feature maps from all previous layers as inputs, and transmits its own feature maps to all subsequent layers, which encourages feature reuse and constructs direct connections among all layers. In addition, compared with ResNet, DenseNet has fewer parameters, and DenseNet with only 0.8 M training parameters can obtain the performance of a 1001-layer ResNet with 10.2 M parameters.

Inspired by the transfer learning strategy and the dense connection approach, we designed a novel end-to-end UNet-based deep convolutional neural network called TL-DenseUNet for the semantic segmentation of remote sensing images with insufficient labeled data and imbalanced data classes. TL-DenseUNet focuses on two aspects. First, it uses transferring DenseNet-121, which is pretrained on ImageNet images (1000 classes), to extract the multiscale semantic features of ground objects from remote sensing images. The transferring parameters can provide prior knowledge to accurately identify multiobject of remote sensing images without overfitting. Second, dense connections are used in the decoder subnetwork to fuse the multiscale semantic features, which can enhance feature reuse and information flow. Our main contributions are as follows:

(1) A UNet-based deep convolutional neural network is proposed in this paper, which performs much better in segmenting multiobject of remote sensing images with insufficient labeled data and imbalanced data classes.

(2) The transferring DenseNet-121 pre-trained on ImageNet is firstly applied in the encoder subnetwork, which plays a guiding role in multiscale feature extraction of remote sensing images.

(3) A novel multiscale fusion module with dense connection is designed in the decoder subnetwork, which can effectively fuse the multiscale semantic features and enhance the recognition power of ground objects from remote sensing images.

The remainder of this article is organized as follows: related works will be discussed in Section II. In Section III we will introduce the details of the proposed TL-DenseUNet. Section IV describes the experimental data and results. Section V presents the discussion. Finally, we will summarize our work in Section VI.

II. RELATED WORKS

In this section, we briefly describe the modern structure of semantic segmentation and two deep learning techniques: transfer learning and dense connection.

A. MODERN STRUCTURE FOR SEMANTIC SEGMENTATION

With breakthroughs in the computational power of graphic processing units (GPUs) and the development of big data, considerable progress in deep convolutional neural networks has occurred in recent years. Semantic segmentation is a successful application of this approach and has been utilized

to solve pixelwise classification problems. Long *et al.* [14] proposed a semantic segmentation technique that replaced the fully connected layers with convolutional layers to enable end-to-end training, and utilized deconvolution [32] layers to predict high-resolution masks from coarse feature maps. In addition, to strengthen the segmentation performance, skip connections between pooling [33] layers were used to fuse the semantic features and appearance features (FCN-8s, FCN-16s and FCN-32s) obtained by the network. The FCNs combined the features from the final three layers (FCN-8s), which made it similar to an incomplete encoder-decoder structure. UNet [34] used a symmetric and complete encoder-decoder structure for biomedical image segmentation, including a contracting path and a similar expanding path in which the pooling layers were replaced by upsampling layers. For precise location, high-resolution features from each layer in the contracting path were combined with upsampled outputs from the corresponding expanding path through a long skip connection. This elegant architecture yielded outstanding performance with very few images. SegNet [35], which was proposed by Vijay *et al.*,officially showed a typical deep convolutional encoder-decoder structure. The encoder was responsible for object classification, and the corresponding decoder reconstructed the encoding features to the same size as the original input. In particular, the decoder used the pooling indices memorized from the corresponding encoder to perform upsampling, which produced a sparse feature map and could be trained effectively. The encoder-decoder structure is the most popular structure for semantic segmentation. RefineNet [36] and global convolutional networks (GCNs) [37], which are based on this structure, have both achieved state-of-the-art performances.

B. TRANSFER LEARNING

Traditional deep convolutional neural networks require large amounts of labeled data for training to achieve optimal performance. The idea of transfer learning is to apply the knowledge learned from a related source task with large amounts of training data to a target task with comparatively insufficient training data in a certain way [38], which helps gradient propagation during training, and reduces the limitations of the data on network performance. Creating labeled data is expensive, so optimally leveraging an existing dataset is key [39]. Some low-level features, such as the edges and shapes of objects, are relevant and can be shared by transferring parameters, allowing the model need not to learn from scratch such as with ordinary networks. Hence, extracting abstract and sophisticated high-level features is the optimization goal of the target task. The most common strategy for transfer learning is to fine-tune a pretrained network model on a target dataset [40]. Girshick *et al.* [25] pretrained a convolutional neural network on ImageNet and then fine-tuned all the network parameters for the target task(detection) where data are insufficient. Long *et al.* [41] fine-tuned only the parameters of the last few layers and confirmed that specific features from these layers tailored to an original task cannot effectively bridge the domain discrepancy. Sharif Razavian *et al.* [10] used a pretrained model to extract features with a support vector machine (SVM) to solve different classification tasks.

C. DENSE CONNECTION

Recent studies have demonstrated the importance of using features from shallow layers to directly optimize features from deep layers, especially for very deep convolutional neural networks. ResNet [30] establishes an identity connection by adding an additional path beyond the normal path between two neurons, which allows the gradient to propagate directly through the bypass. Moreover, the identity function and residual function are combined through summation. The residual connection between the front layer and the back layer effectively alleviates the problems of the vanishing gradients and model degradation as the number of network layers increases. DenseNet [31] offers another typical approach called dense connection, which mainly consists of dense blocks. Figure 1 shows a basic nonlinear transformation module (a) in a dense block and a typical dense block (b). In contrast to residual connections, dense connections combine features by concatenation. In each dense block, all the layers are directly connected to each other, ensuring maximum information flow between layers (b). Therefore, the l^{th} feature x_l receives the output of the preceding $l-1$ features $x_0, \ldots, x_{l-1}$ as input:

$$x_l = H_l([x_0, \ldots, x_{l-1}]) \tag{1}$$

where $[x_0, \ldots, x_{l-1}]$ refers to the concatenation of the output from layers $0, \ldots, l-1$ and H_l is defined as a nonlinear transformation module (a). The multiple inputs and the output of H_l are combined by concatenation.

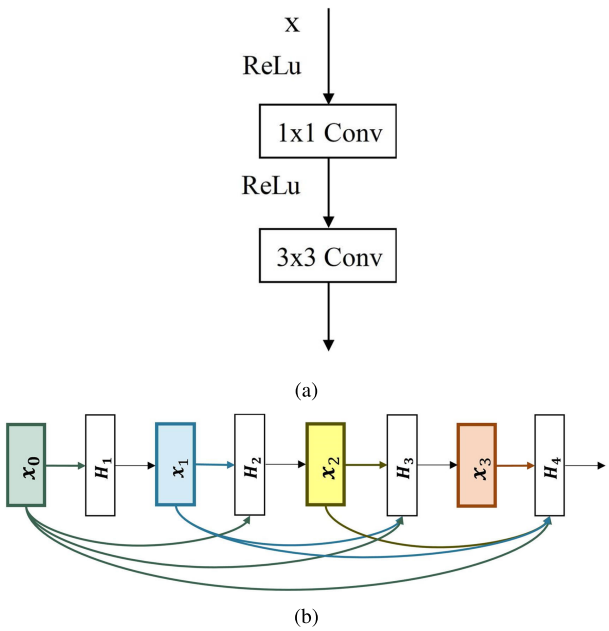


FIGURE 1. A basic nonlinear transformation module (a) in a dense block and a typical dense block (b) in densely connected convolutional networks.

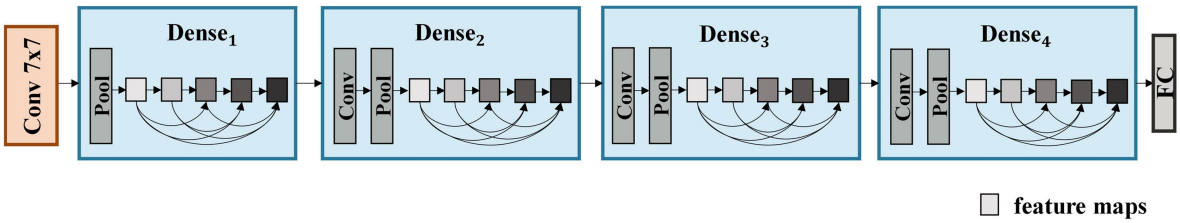


FIGURE 2. The structure of DenseNet-121.

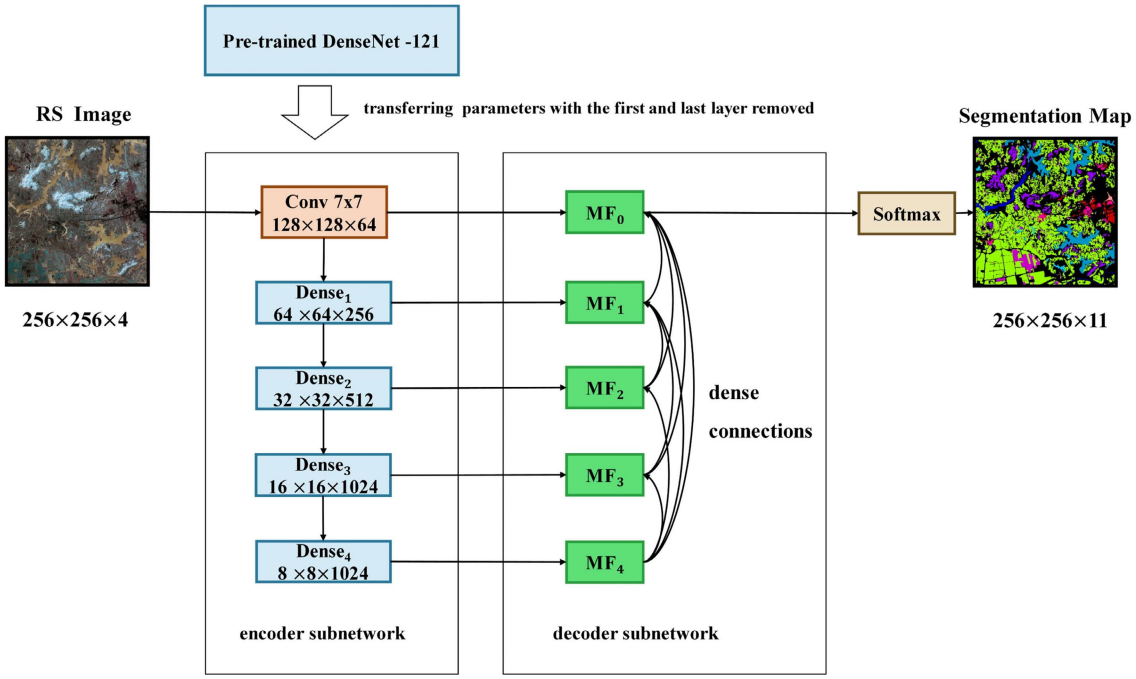


FIGURE 3. Overview of TL-DenseUNet for the semantic segmentation of remote sensing images.

An $L-layer$ model produces $\frac{L(L+1)}{2}$ connections instead of only L, as in the traditional structure, which strengthens the information flow among layers. In addition, feature reuse means that DenseNet requires fewer parameters than a traditional convolutional neural network, because there is no need to learn the redundant features obtained from earlier layers via concatenation. The direct connections among all the layers improve gradient propagation during training and alleviate the vanishing gradient problem. Additionally, each layer can obtain the gradient directly from the loss function and the original input, which represents a kind of implicit deep supervision and helps train the deeper network.

Figure 2 shows a classic version of DenseNet with dense connections—DenseNet121, which mainly consists of four parts ($Dense_1$, $Dense_2$, $Dense_3$ and $Dense_4$). Each part has a 2×2 average pooling operation to reduce the resolution of the feature maps. Moreover, there is also an 1×1 convolutional operation in the last three Dense parts to reduce the number of feature maps. Inside the Dense parts, dense connections are constructed from any layer to all subsequent layers after the pool layers.

III. METHODS

This section presents the proposed TL-DenseUNet. First, the network architecture of TL-DenseUNet is introduced in Section III-A. Then, the transfer learning strategy for multiscale feature extraction in TL-DenseUNet is described in Section III-B. Finally, the multiscale fusion module we designed is exhibited in detail in Section III-C.

A. THE NETWORK ARCHITECTURE OF TL-DenseUNet

The TL-DenseUNet proposed in this paper for remote sensing image semantic segmentation is based on UNet [34], which is an end-to-end network and has two symmetric encoder-decoder subnetworks. The input of TL-DenseUNet is a remote sensing image and the output is a categorical segmentation map.

As shown in Figure 3, the transferring DenseNet-121 is employed in the encoder subnetwork to extract the multiscale features of remote sensing images. The detailed transfer learning strategy can be found in Section III-B. Note that, the transferring DenseNet-121 directly removes the fully

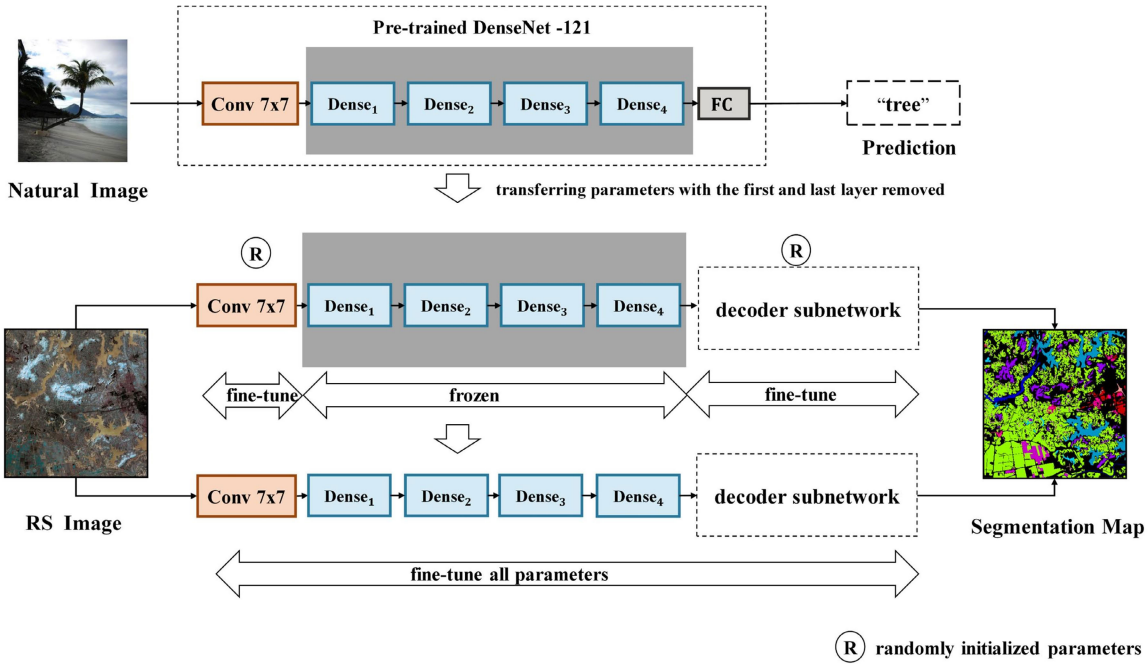


FIGURE 4. Transfer learning in TL-DenseUNet.

connected layer to ensure end-to-end training and avoid the loss of spatial information.

The decoder subnetwork is responsible for fusing the object features extracted by the encoder subnetwork, which mainly consists of five multiscale fusion (MF) modules. Skip connection is utilized between each MF module and the corresponding Dense module in the encoder subnetwork. To improve information flow and feature reuse, dense connections among the MF modules are designed to ensure the precise fusion of multiscale semantic features from different levels. Details about the MF module can be found in Section III-C.

At last, a softmax function is used to calculate the classification probability distribution and derive the semantic segmentation map.

B. TRANSFER LEARNING STRATEGY IN TL-DenseUNet

As mentioned previously, the transfer learning strategy is leveraged to train our deep convolutional neural network. As presented in the top of Figure 4, DenseNet-121 is pre-trained on the ImageNet dataset with three bands (R, G, B). To make it fit our target segmentation task, we remove the last fully connected (FC) layer and treat the rest of DenseNet-121 as a multiscale feature extractor. Note that to transfer the model to segment n-band remote sensing images, we adjust the channels of the first convolution kernel in the original model from three to n. The four Dense modules of TL-DenseUNet are initialized with transferring parameters, and the Conv 7×7 kernel and the decoder subnetwork are randomly initialized by an initialization function.

The transfer learning strategy mainly includes two stages: fine-tuning part of the network and fine-tuning all the network. First, as shown in the middle of Figure 4, we freeze the transferring parameters, which means that these parameters would not be updated in this stage, and fine-tune the randomly initialized parameters in the Conv 7×7 and the decoder subnetwork using target training data for some epochs. Then, as shown in the bottom of Figure 4, we use target training data to fine-tune the entire network. All the parameters of the network are trained together to achieve more excellent performance on the target segmentation task.

C. MULTISCALE FUSION MODULE

The multiscale fusion (MF) module we designed aims at fusing the local information extracted from the corresponding encoder layer and the semantic information derived from the previous decoder layers. As shown in Figure 5 (a), the decoder subnetwork consists of five MF modules. In order to strengthen multiscale feature reuse and improve information flow, dense connections are introduced among the MF modules. It is the first time that dense connections are applied in the decoder of UNet-like network.

As shown in Figure 5 (b), each MF module consists of two Conv 1×1, one Conv 3×3, two concatenation operations and multiple UpSample operations. The Conv 1×1 is applied to reduce the number of input feature maps, which has proved its effectiveness in dimension reduction [42]. The Conv 3×3 is introduced to capture the context information. Note that each Conv in the MF module represents the three consecutive operations: batch normalization (BN) [43], convolution (Conv) and a rectified linear unit (ReLU) [44].

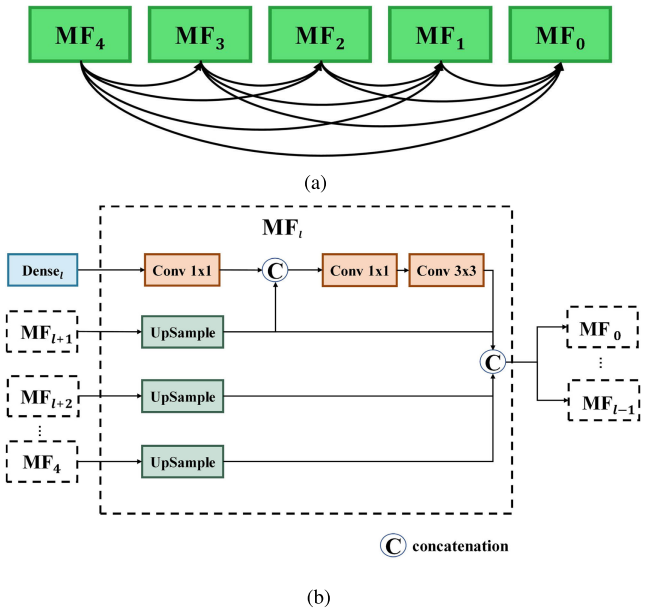


FIGURE 5. Dense connections among MF modules (a) and the architecture of each MF module (b).

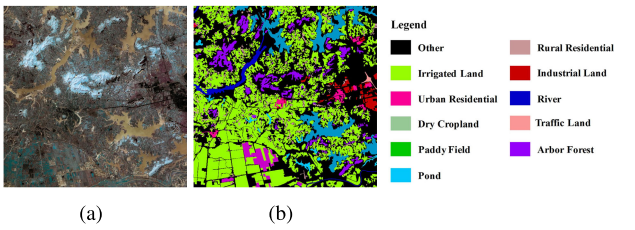


FIGURE 6. Example of a remote sensing image used in the test. (a) is a four-band image (R, G, B, NIR) and (b) is the corresponding ground truth.

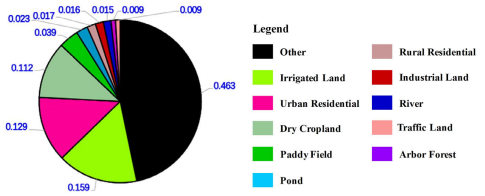


FIGURE 7. Proportional distribution of each class in the dataset.

The two concatenation operations are responsible for multiscale feature fusion. Following the structure of UNet, skip connection is utilized in the first concatenation operation to fuse the output from the corresponding Dense module with the output from the previous MF module which should first undergo an UpSample operation. Then, following the idea of dense connection, the second concatenation operation is utilized to fuse the multiscale features from all preceding MF modules. However, the concatenation operation does not work when the size of the feature map changes. Hence, we must make sure that the feature maps we want to fuse have the same spatial resolution. So, the feature maps from all preceding MF modules should first undergo an UpSample operation to enlarge the low-resolution feature maps to the largest size of the feature maps we want to fuse. Finally, these feature maps are concatenated together as the output of the current MF module and transferred to all subsequent MF modules. Note that, the MF_4 module has only one input from the encoder subnetwork, therefore it does not have any fusion parts.

IV. EXPERIMENTS AND EVALUATION

In this section, we briefly exhibit the test dataset and implementation details. Then, we evaluate the performance of the proposed TL-DenseUNet model in semantic segmentation of remote sensing images.

A. DATASET

The dataset comes from the Remote Sensing Image Sparse Representation and Intelligent Analysis competition (http://rscup.bjxintong.com.cn/) held by the Information Science Department of the National Natural Science Foundation of China (NSFC) in 2019. All the images were taken by Gaofen-2 (GF2). Each remote sensing image has at least 7200 × 6800 pixels, which take a value between 0 and 255. The image consists of four spectral bands (R, G, B, NIR) with a spatial resolution of 4*m* per pixel. Among them, ten ground truths could be obtained and each ground truth was segmented into eleven pixel-level classes: paddy field, irrigated land, dry cropland, arbor forest, traffic land, industrial land, rural residential, urban residential, river, pond, and other (background). Figure 6 (a) and (b) show one of the images and the corresponding ground truth. The proportional distribution of each class in the dataset is shown in Figure 7. We can see that the distribution of data classes is imbalanced, and some classes of data, such as arbor forests, are rare. It is obvious that the imbalanced and insufficient labeled data makes the semantic segmentation task very challenging. We divided ten images into a training set of seven images, a validation set of one image and a testing set of two images. During training and validation, the images are randomly clipped into 256 × 256 overlapping patches by using the sliding window algorithm with a stride of 128 pixels. The final dataset includes 20020 samples for training, 2860 samples for validation, and 5720 samples for testing. Common data augmentation methods were also used to avoid overfitting and optimize training. In the training stage, the images may be first preprocessed with one of the following operations or a combined operation: flip (horizontal or vertical) and add noise.

B. IMPLEMENTATION DETAILS

We implemented TL-DenseUNet in Keras [45] with TensorFlow [46] as the backend. Note that the remote sensing image dataset consists of four bands, so we adjust the channels of the first convolutional kernel in the pretrained DenseNet-121 from three to four. The experiment was performed on a Linux platform with an NVIDIA P100 GPU (16 GB RAM).

TABLE 1. Quantitative scores obtained from the semantic segmentation of remote sensing images. "P":"precision"; "R":"recall"; "F":"F1 score"; "I":"IoU".

	Method	Paddy Field	Irrigated Land	Dry Crop land	Arbor Forest	Industrial Land	Urban Residential	Rural Residential	River	Pond	Traffic Land	Other
P(%)	UNet	35.67	62.56	3.68	**96.71**	13.39	55.06	9.51	31.96	47.79	22.58	61.69
	SegNet	15.31	71.34	4.63	92.92	34.45	68.65	13.59	50.09	54.96	26.69	62.32
	DeepResUNet	1.09	75.46	5.05	89.33	43.53	60.46	5.79	56.73	59.09	27.78	61.74
	RefineNet	88.68	**83.83**	8.19	79.39	18.01	**80.41**	19.29	56.87	60.18	41.11	68.48
	TL-DenseUNet	**90.43**	82.68	**12.81**	90.78	**52.48**	57.42	**35.32**	**79.12**	**69.77**	**45.57**	**70.52**
R(%)	UNet	1.69	60.36	3.49	58.93	1.06	62.32	1.77	23.79	**61.54**	0.01	74.83
	SegNet	1.42	48.39	4.77	57.82	46.02	63.11	14.45	26.53	58.49	6.93	83.09
	DeepResUNet	0.05	25.65	7.65	70.60	38.24	66.07	1.59	36.47	60.03	37.21	**87.05**
	RefineNet	0.03	74.71	25.39	3.56	**50.46**	45.01	**37.17**	47.19	42.37	45.01	80.53
	TL-DenseUNet	**22.76**	**80.11**	**32.60**	**77.49**	47.95	**78.77**	35.19	**49.55**	61.04	**60.70**	83.12
F(%)	UNet	3.24	61.43	3.58	73.23	1.96	58.45	2.12	23.97	37.30	0.01	67.62
	SegNet	2.59	57.67	4.70	71.28	39.40	65.76	9.27	27.36	43.07	11.01	70.98
	DeepResUNet	0.10	38.29	6.08	78.87	39.59	62.89	2.47	36.39	43.83	31.81	71.12
	RefineNet	0.05	79.01	15.39	6.82	45.73	67.39	22.49	52.95	55.62	42.10	73.66
	TL-DenseUNet	**36.37**	**81.37**	**18.39**	**83.61**	**50.40**	**73.52**	**24.18**	**60.64**	**57.94**	**52.06**	**74.99**
I(%)	UNet	1.64	46.34	1.57	63.77	1.12	45.61	1.06	15.45	23.81	0.01	55.08
	SegNet	1.31	40.52	2.41	55.38	33.32	49.67	5.03	17.67	27.97	5.82	57.03
	DeepResUNet	0.05	23.68	3.14	65.11	33.17	47.56	1.26	25.67	28.44	21.91	57.13
	RefineNet	0.03	69.29	13.60	6.53	32.65	52.79	17.31	37.61	39.72	28.41	59.87
	TL-DenseUNet	**21.23**	**69.60**	**15.31**	**70.84**	**35.35**	**61.41**	**18.80**	**43.56**	**40.89**	**35.19**	**61.68**

The Adaptive Moment Estimation (Adam) [47] algorithm was used as the optimization algorithm to minimize training loss and update model parameters. During training, we first froze the parameters of DenseNet-121 and trained them for ten epochs. The initial learning rate was set to 0.0003. Then, the entire model was trained for fifty epochs with an initial learning rate of 0.0001 and a weight decay of 0.00001. Due to the limit of the GPU memory, the batch size during training was set to eight in the experiment.

To quantitatively evaluate the performance of the proposed TL-DenseUNet in segmenting remote sensing images, seven traditional metrics were applied: the precision, recall, F1 score, IoU, overall accuracy (OA), kappa coefficient(Kappa) and MIoU.

C. RESULTS AND COMPARISONS

To evaluate the effectiveness of TL-DenseUNet, we selected some state-of-the-art models for comparison: UNet [34], SegNet [35], DeepResUNet [22], and RefineNet [36]. Note that all models were tested with all test images in the same experimental environment. Note that, as described above, we fine-tuned the randomly initialized parameters in TL-DenseUNet's Conv 7×7 and decoder subnetwork for ten epochs. Hence, to be fair, all the other models mentioned above were trained for ten more epochs than TL-DenseUNet. The quantitative scores obtained are shown in Table 1, and the best values of each metric are shown in bold.

As shown in Table 1, TL-DenseUNet outperformed other advanced semantic segmentation models largely in terms of most metrics. Among these models, UNet displayed the worst performance, followed by SegNet, implying that simple models have difficulty segmenting remote sensing images with insufficient and imbalanced labeled data. DeepResUNet performed better than the above two methods, but it still had very low metrics, such as the IoU for the paddy field class. The performance of RefineNet was second to that of TL-DenseUNet; however, it performed very poorly for some classes with relatively limited data. Benefiting from the transferring DenseNet-121 and the MF module, TL-DenseUNet achieved relatively satisfactory performances in both precision and recall, and obtained the best F1 score and IoU. For the classes with more data, such as irrigated land, the F1 score of TL-DenseUNet (81.37%) exceeded that of DeepResUNet (38.29%), UNet (61.43%), RefineNet (79.01%), and SegNet (57.67%), which demonstrated the effectiveness of the MF module for enhancing the recognition ability of ground objects in remote sensing images. For the performance of segmenting classes with limited data, such as arbor forest, TL-DenseUNet's F1 score was improved by at least 4.74% and IoU was improved by at least 5.73%, indicating that the transferring DenseNet-121 improved multiscale feature extraction from remote sensing images. These findings demonstrate the superiority of TL-DenseUNet in the semantic segmentation of remote sensing images with insufficient and imbalanced labeled data.

Table 2 reports the OA, kappa coefficient and MIoU obtained by the five models. As can be seen from the results in Table 2, TL-DenseUNet was still the best among all models. UNet and SegNet had the worst performances, which further confirmed that simple models are not suitable for the segmentation of remote sensing images with insufficient and imbalanced labeled data. DeepResUNet and RefineNet achieved better results than the above two methods, but the results were still not satisfactory. DeepResUNet was originally proposed to extract buildings from remote sensing images, which made it difficult to segment complex remote

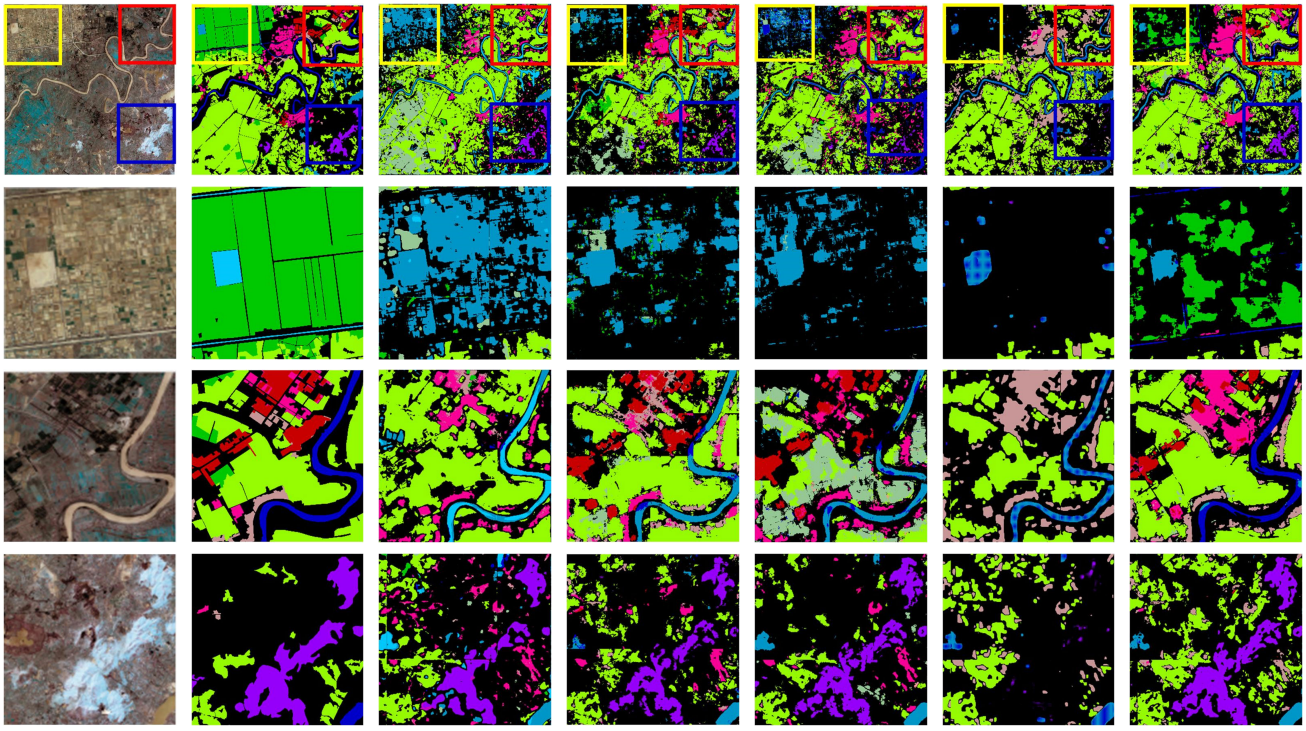


FIGURE 8. Visual comparisons between TL-DenseUNet (ours) and other models. The first row shows the overall results of the test image and the last three rows show three randomly selected areas from the overall results. (a) Image. (b) Ground Truth. (c) UNet. (d) SegNet. (e) DeepResUNet. (f) RefineNet. (g) TL-DenseUNet.

sensing images. TL-DenseUNet achieved the best OA, kappa coefficient and MIoU, indicating that it has the best performance in segmenting remote sensing images with insufficient and imbalanced labeled data.

Figure 8 shows the overall visual segmentation results of the five models for one test image. As can be seen from these figures, SegNet, UNet, DeepResUNet, and RefineNet had difficulties in segmenting ground objects with limited data such as ponds, rivers, and arbor forests. These models rarely accurately identified the paddy field which has a small-sized pond located in it, and were easy to misclassify paddy fields as ponds because they were close to each other. Moreover, these models also often misclassified rivers as ponds (shown in the third row of Figure 8) because these two objects had high similarity. In contrast, TL-DenseUNet performed relatively better than the other models. With the proposed methods, major parts of the paddy fields can be extracted, and paddy fields were rarely misclassified as ponds. Moreover, the main extraction of arbor forests and the edge extraction of rivers were also improved, indicating that the transferring DenseNet-121 and the MF module help achieve better performance than the other state-of-the-art methods.

For further comparison, the semantic segmentation results of some of the classes in the test images are shown in Figure 9. As can be seen, TL-DenseUNet achieved the best performance in extracting multiobject from remote sensing images, while the other models had more false positives (red) and false negatives (green) in the semantic segmentation of each ground object. DeepResUNet displayed the worst performance because too many false negatives (green) appeared in the semantic segmentation result for irrigated land. SegNet and UNet improved the segmentation quality of irrigated land, but still generated more incomplete and inaccurate segmentation results for arbor forests than did the proposed method; moreover, they also did not distinguish well between rivers and ponds. RefineNet yielded relatively good performance in segmenting irrigated land, but most false negatives (green) appeared in the segmentation of arbor forests, indicating that many arbor forests were not accurately identified. It was clear that these four models did not achieve satisfactory results for the segmentation of urban residential. TL-DenseUNet not only had fewer false positives and false negatives in the segmentation of rivers, arbor forests

TABLE 2. OA, Kappa and MIoU obtained from the semantic segmentation of remote sensing images.

Method	OA(%)	Kappa	MIoU(%)
UNet	61.41	0.3992	23.22
SegNet	61.76	0.4038	26.92
DeepResUNet	62.05	0.4113	27.92
RefineNet	67.54	0.4917	32.53
TL-DenseUNet	**72.01**	**0.5669**	**43.08**

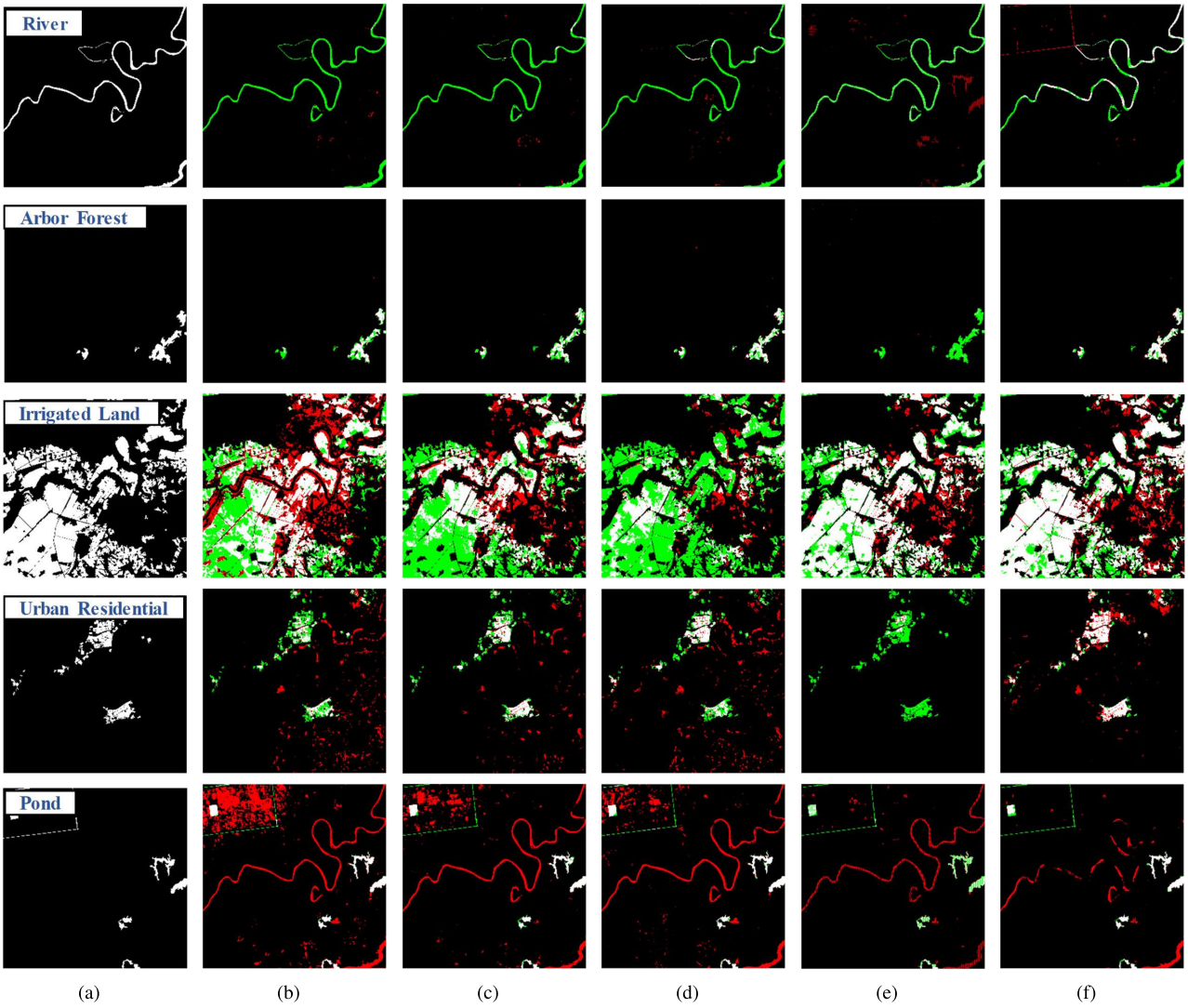


FIGURE 9. Visual comparisons of the segmentation results of the five models. Among them, the white, red and green areas represent true positive, false positive and false negative predictions, respectively (a) Ground Truth. (b) UNet. (c) SegNet. (d) DeepResUNet. (e) RefineNet. (f) TL-DenseUNet.

and ponds, which had limited data, but it was also able to extract more accurate irrigated land and urban residential from remote sensing images. These facts further verify that the transferring DenseNet-121 used in the encoder subnetwork and the MF module designed in the decoder subnetwork help TL-DenseUNet perform better than the other state-of-the-art models.

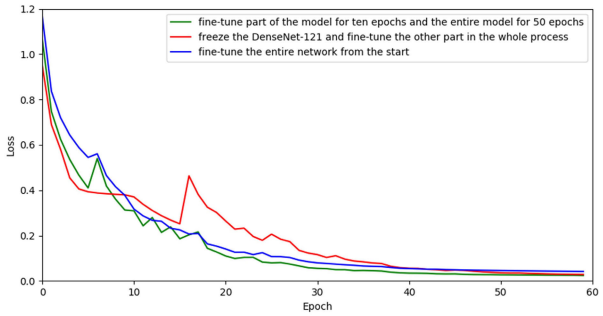
D. COMPARISONS OF DIFFERENT TRANSFER LEARNING STRATEGIES

Figure 10 shows the loss and accuracy curves for the different transfer learning strategies on the training of TL-DenseUNet. As shown in Figure 10, freezing DenseNet-121 throughout the training process and fine-tuning other parts of the model (Strategy-1) yields relatively high loss and low accuracy. This result demonstrates that not updating some parameters all the time affects the ability of the model to extract multilevel semantic features from target data. Fine-tuning the entire network from the start (Strategy-2) achieves better performance than the first one. The loss curve for the strategy of freezing the parameters of DenseNet-121 for ten epochs and fine-tuning the entire model for fifty epochs (Strategy-3) displays lower value than that for the other two strategies. Moreover, as shown in Table 3, Strategy-3 achieves the best OA, kappa coefficient and MIoU of the three strategies. These facts highlight the effectiveness of the transfer learning strategy used in our experiment.

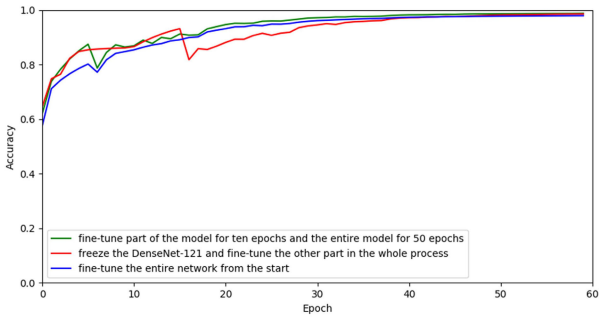
V. DISCUSSION

A. EFFECTS OF THE TRANSFERRING DenseNet-121 AND DENSE CONNECTION

The superior performance of TL-DenseUNet was mainly related to two strategies: the transferring DenseNet-121 (TFD) used in the encoder subnetwork and the dense



(a)



(b)

FIGURE 10. Loss (a) and accuracy (b) curves of each epoch obtained by TL-DenseUNet while using different transfer learning strategies.

TABLE 3. OA, Kappa and MIoU obtained by TL-DenseUNet while using different transfer learning strategies.

Method	OA(%)	Kappa	MIoU(%)
Strategy-1	70.93	0.5512	41.09
Strategy-2	71.79	0.5607	42.11
Strategy-3	**72.01**	**0.5669**	**43.08**

connection (DC) used in the MF module. Benefiting from these two strategies, all the evaluation metrics were greatly improved compared to those of traditional methods. To demonstrate that both strategies can improve the semantic segmentation performance of remote sensing images, we compared the accuracies among the different variants of TL-DenseUNet. Note that we used TL-DenseUNet without the transferring DenseNet-121 (TFD) and dense connection (DC) as our baseline. The experimental setup was the same as before. All the quantitative evaluation metrics are shown in Table 4.

As shown in Table 4, when the transferring DenseNet-121 (TFD) was added to the designed model for training, the performance was significantly improved. The OA, kappa coefficient and MIoU have been improved by 4.22%, 0.0686 and 7.12%, implying that reusing the pretrained parameters from the natural image can greatly improve the model performance, even though the bands of the target remote sensing image are different. It is probably because that compared with training from scratch, using the pretrained parameters

TABLE 4. Comparison of the accuracies among the different variants of TL-DenseUNet, and the best values are in bold.

Method	OA(%)	Kappa	MIoU(%)
Baseline	65.03	0.4585	29.11
Only TFD	69.25	0.5271	36.23
Only DC	66.91	0.4791	31.26
TL-DenseUNet	**72.01**	**0.5669**	**43.08**

to initialize TL-DenseUNet can provide guidance for model convergence. When TL-DenseUNet added only the dense connection (DC) in the MF module, the OA, kappa coefficient and MIoU have been improved by 1.88%, 0.0206 and 2.15%. It is obvious that the reuse of semantic information obtained from the previous decoder layers can guide feature reconstruction of remote sensing images, which ensures the full use of features to generate more accurate segmentation maps. When adding both, the OA, kappa coefficient and MIoU were improved by 6.98%, 0.1084 and 13.97%, which indicated that using the two strategies simultaneously can further improve the model performance.

B. MODEL COMPLEXITY

For further analysis, we compared the number of parameters, training time and inference time with those of UNet, SegNet, DeepResUNet, and RefineNet. The time required to load the pretrained parameters was excluded from the training and inference time. The dataset used in the training stage was the same as that used before, and the size of the image used in the inference stage was 512×512.

TABLE 5. Comparisons of model complexity.

Method	Number of Parameters (M)	Training Time (Seconds/Epoch)	Inference Time (ms/Image)
UNet	31.03	1335	75
SegNet	29.46	1796	111
DeepResUNet	2.79	1815	78
RefineNet	49.25	3075	189
TL-DenseUNet	13.19	2369	170

As shown in Table 5, there are fewer parameters in TL-DenseUNet than in most models, except for DeepResUNet. DeepResUNet has the fewest parameters, because it is a lightweight model. However, its performance was much poorer than that of our model. TL-DenseUNet follows the structure of UNet, but it adopts dense connection in both the encoder and decoder subnetworks, which makes it possible to generate relatively few feature maps in each layer. This is the reason why TL-DenseUNet has fewer parameters. RefineNet has the most parameters and the longest training and inference time, which may be caused by its complex structure. UNet and SegNet require relatively short training and inference time, because they have simple convolution and pooling operations, which leads to simple gradient flow. TL-DenseUNet requires longer time to train and inference than UNet because the dense blocks used in each layer of

the encoder subnetwork require more calculations. Moreover, the dense connections inside TL-DenseUNet lead to the complexity of gradient propagation, which may have a negative effect on model training. It is also an aspect for us to improve in the future.

VI. CONCLUSIONS

In this paper, a novel UNet-based deep convolutional neural network, called TL-DenseUNet, was proposed to segment multiobject from remote sensing images with insufficient labeled data and imbalanced data classes. TL-DenseUNet adopts a transferring DenseNet-121 and multiple MF modules to enhance model performance. Experiments were carried out on a remote sensing image dataset with 11 classes. Both visual and quantitative experimental results demonstrated that the transfer learning strategy can deal with the problem of insufficient and imbalanced samples more effectively, and the MF modules we designed can enhance feature reuse and information flow. Moreover, our work verified that transferring network parameters from three-band natural images to multiband remote sensing images is also effective. However, the overall performance of the proposed TL-DenseUNet was still not so satisfactory. Ground objects with similar spectra, such as rivers and ponds, were prone to be misclassified. In the future, we will explore an unsupervised transfer learning method, which can leverage large amounts of unlabeled remote sensing images and reduce the labeling costs, to achieve more accurate semantic segmentation.

ACKNOWLEDGMENT

The authors would like to thank all reviewers and editors for their comments on this paper. The authors would also like to thank the Information Science Department of the National Natural Science Foundation of China (NSFC) for providing the test dataset.

REFERENCES

[1] Y. Xu, L. Wu, Z. Xie, and Z. Chen, "Building extraction in very high resolution remote sensing imagery using deep learning and guided filters," *Remote Sens.*, vol. 10, no. 1, p. 144, Jan. 2018.

[2] G. Cheng, F. Zhu, S. Xiang, Y. Wang, and C. Pan, "Accurate urban road centerline extraction from VHR imagery via multiscale segmentation and tensor voting," *Neurocomputing*, vol. 205, pp. 407–420, Sep. 2016.

[3] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[4] 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, Nov. 2018.

[5] L. Matikainen and K. Karila, "Segment-based land cover mapping of a suburban area-comparison of high-resolution remotely sensed datasets using classification trees and test field points," *Remote Sens.*, vol. 3, no. 8, pp. 1777–1804, Aug. 2011.

[6] J. V. Solórzano, J. A. Meave, J. A. Gallardo-Cruz, E. J. González, and J. L. Hernández-Stefanoni, "Predicting old-growth tropical forest attributes from very high resolution (VHR)-derived surface metrics," *Int. J. Remote Sens.*, vol. 38, no. 2, pp. 492–513, Jan. 2017.

[7] T. Shi, Q. Xu, Z. Zou, and Z. Shi, "Automatic raft labeling for remote sensing images via dual-scale homogeneous convolutional neural network," *Remote Sens.*, vol. 10, no. 7, p. 1130, Jul. 2018.

[8] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 166–177, Jun. 2019.

[9] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[10] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features Off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 806–813.

[11] P. Liu, X. Liu, M. Liu, Q. Shi, J. Yang, X. Xu, and Y. Zhang, "Building footprint extraction from high-resolution images via spatial residual inception convolutional neural network," *Remote Sens.*, vol. 11, no. 7, p. 830, Apr. 2019.

[12] Q. Shi, X. Liu, and X. Li, "Road detection from remote sensing images by generative adversarial networks," *IEEE Access*, vol. 6, pp. 25486–25494, 2018.

[13] M. Lan, Y. Zhang, L. Zhang, and B. Du, "Global context based automatic road segmentation via dilated convolutional neural network," *Inf. Sci.*, vol. 535, pp. 156–171, Oct. 2020.

[14] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[15] J. Sherrah, "Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery," 2016, *arXiv:1606.02585*. [Online]. Available: http://arxiv.org/abs/1606.02585

[16] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, Atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[17] K. Bittner, F. Adam, S. Cui, M. Korner, and P. Reinartz, "Building footprint extraction from VHR remote sensing images combined with normalized DSMs using fused fully convolutional networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 8, pp. 2615–2629, Aug. 2018.

[18] G. Chen, X. Zhang, Q. Wang, F. Dai, Y. Gong, and K. Zhu, "Symmetrical dense-shortcut deep fully convolutional networks for semantic segmentation of very-high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1633–1644, May 2018.

[19] S. Chakraborty, V. Balasubramanian, Q. Sun, S. Panchanathan, and J. Ye, "Active batch selection via convex relaxations with guaranteed solution bounds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 1945–1958, Oct. 2015.

[20] J. Ding, B. Chen, H. Liu, and M. Huang, "Convolutional neural network with data augmentation for SAR target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 3, pp. 364–368, Mar. 2016.

[21] B. Pan, Z. Shi, and X. Xu, "R-VCANet: A new deep-learning-based hyperspectral image classification method," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 5, pp. 1975–1986, May 2017.

[22] Y. Yi, Z. Zhang, W. Zhang, C. Zhang, W. Li, and T. Zhao, "Semantic segmentation of urban buildings from VHR remote sensing imagery using a deep convolutional neural network," *Remote Sens.*, vol. 11, no. 15, p. 1774, Jul. 2019.

[23] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.

[24] S. Akcay, M. E. Kundegorski, M. Devereux, and T. P. Breckon, "Transfer learning using convolutional neural networks for object classification within X-ray baggage security imagery," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 1057–1061.

[25] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[26] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1717–1724.

[27] B. Pan, Z. Shi, X. Xu, T. Shi, N. Zhang, and X. Zhu, "CoinNet: Copy initialization network for multispectral imagery semantic segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 5, pp. 816–820, May 2019.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[29] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning Earth observation classification using ImageNet pretrained networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, Jan. 2016.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[31] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[32] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2018–2025.
[33] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[34] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[35] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[36] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1925–1934.
[37] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters-improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4353–4361.
[38] W. Zhang, Y. Zhu, and Q. Fu, "Semi-supervised deep transfer learning-based on adversarial feature learning for label limited SAR target recognition," *IEEE Access*, vol. 7, pp. 152412–152420, 2019.
[39] T. Panboonyuen, K. Jitkajornwanich, S. Lawawirojwong, P. Srestasathiern, and P. Vateekul, "Semantic segmentation on remotely sensed images using an enhanced global convolutional network with channel attention and domain specific transfer learning," *Remote Sens.*, vol. 11, no. 1, p. 83, Jan. 2019.
[40] Y. Guo, H. Shi, A. Kumar, K. Grauman, T. Rosing, and R. Feris, "SpotTune: Transfer learning through adaptive fine-tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4805–4814.
[41] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," 2015, *arXiv:1502.02791*. [Online]. Available: http://arxiv.org/abs/1502.02791
[42] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[43] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167
[44] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.
[45] N. K. Manaswi, N. K. Manaswi, and S. John, *Deep Learning with Applications Using Python*. Cham, Switzerland: Springer, 2018.
[46] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Oper. Syst. Design Implementation*, 2016, pp. 265–283.
[47] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

BINGE CUI received the B.Sc., M.Sc., and Ph.D. degrees in computer science from Harbin Engineering University, Harbin, China, in 2000, 2003, and 2006, respectively.

In 2006, he joined the College of Computer Science and Engineering, Shandong University of Science and Technology. From 2010 to 2011, he was a Visiting Scholar with the Department of Information System, City University of Hong Kong. From 2012 to 2014, he was a Postdoctoral Researcher with the First Institute of Oceanography, State Oceanic Administration, China. He is currently a Professor. His research interests include hyperspectral image classification and remote sensing image analysis with deep learning.

XIN CHEN received the bachelor's degree from Jiangsu Ocean University, Lianyungang, China, in 2018. He is currently pursuing the master's degree with the College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, China.

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

YAN LU received the B.Sc. and M.Sc. degrees in computer science from Yanshan University, Qinhuangdao, China, in 1998 and 2000, respectively, and the Ph.D. degree in computer science from Fudan University, Shanghai, China, in 2003. From 2003 to 2005, she was a Postdoctoral Researcher in computer science and technology with the Harbin Institute of Technology. In 2005, she joined the College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao, China. She is currently an Associate Professor. Her research interests include hyperspectral image classification and object detection.