MSNet: multispectral semantic segmentation network for remote sensing images

Chongxin Tao\(^{a,b,}\), Yizhuo Meng\(^{c}\), Junjie Li\(^{a}\), Beibei Yang\(^{a}\), Fengmin Hu\(^{a}\), Yuanxi Li\(^{a}\), Changlu Cui\(^{a}\) and Wen Zhang\(^{a}\)

\(^{a}\)School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China; \(^{b}\)College of Public Administration, Xinjiang Agricultural University, Urumqi, China; \(^{c}\)Department of Computer Science, University of Southern California, Los Angeles, California, USA

**ABSTRACT**

In the research of automatic interpretation of remote sensing images, semantic segmentation based on deep convolutional neural networks has been rapidly developed and applied, and the feature segmentation accuracy and network model generalization ability have been gradually improved. However, most of the network designs are mainly oriented to the three visible RGB bands of remote sensing images, aiming to be able to directly borrow the mature natural image semantic segmentation networks and pre-trained models, but simultaneously causing the waste and loss of spectral information in the invisible light bands such as near-infrared (NIR) of remote sensing images. Combining the advantages of multispectral data in distinguishing typical features such as water and vegetation, we propose a novel deep neural network structure called the multispectral semantic segmentation network (MSNet) for semantic segmentation of multi-classified feature scenes. The multispectral remote sensing image bands are split into two groups, visible and invisible, and ResNet-50 is used for feature extraction in both coding stages, and cascaded upsampling is used to recover feature map resolution in the decoding stage, and the multi-scale image features and spectral features from the upsampling process are fused layer by layer using the feature pyramid structure to finally obtain semantic segmentation results. The training and validation results on two publicly available datasets show that MSNet has competitive performance. The code is available: [https://github.com/taochx/MSNet](https://github.com/taochx/MSNet).

**1. Introduction**

Deep learning algorithms learn representative and distinguishing features hierarchically from data and have been introduced into the field of remote sensing and developed rapidly. Information such as spectra and textures of remote sensing images is fed into convolutional neural network models as the underlying features for pixel-based semantic segmentation to derive feature classification information. Compared with the traditional optical remote sensing image segmentation, the deep feature-based segmentation method can utilize neural networks to implicitly establish pixel-to-semantic mapping relationships (Schuegraf and Bittner 2019), and the network automatically learns to extract the target features of remote sensing images to complete the whole segmentation process without introducing manual feature intervention while improving the accuracy of the results and the generalization ability of the model (Li et al. 2021).

Different from digital pictures, the high complexity of backgrounds and shadows in remote sensing images, scale multilateralism, overly dispersed targets, and gradually increasing high resolution all increase the difficulty of segmentation tasks (Liu et al. 2021). Efficient and intelligent remote sensing image segmentation by deep feature mining has become a research hotspot (Liu et al. 2018), but currently deep learning algorithms cannot be directly applied to many tasks such as remote sensing image interpretation or data processing, the most notable reason being a large number of bands in remote sensing images.

Multispectral remote sensing images contain data of at least four bands or more. The conventional method is to reflect the spectral feature information of different ground objects through different combinations or calculations between the visible and invisible light bands and use them for classification and interpretation. However, most of the processing methods based on deep learning algorithms follow the design method of semantic segmentation of natural images, mainly in the three bands of visible light, RGB, and can directly apply classical semantic segmentation networks for digital pictures: DeepLab (Chen et al. 2018) or UNet (Ronneberger, Fischer, and Brox 2015), and use pre-
trained models: ResNet (He et al. 2016) as backbone network. This network structure is not suitable for multispectral remote sensing images containing non-conventional data such as NIR and digital surface model (DSM) (Yuan, Shi, and Gu 2021). The commonly adopted solutions are mainly to increase the number of channels in the network input layer and to use image fusion techniques (Carvalho et al. 2021). Although the scheme of expanding the number of input channels is effective, it also changes the structure of the pre-trained model, causing its weight value to fail, and the final multispectral data is not significantly improved compared to the RGB model training results. Image fusion technology has gradually matured. The network adopts a dual-branch structure to integrate image and spectral information to improve segmentation accuracy. Image feature extraction can fully call the structure and weight values of the pre-trained model while taking into account the extraction of spectral features and geometric features of remote sensing images (Ha et al. 2017; Hazirbas et al. 2017; Huang et al. 2020; Song, Lichtenberg, and Xiao 2015; Sun, Zuo, and Liu 2019; Yuan, Shi, and Gu 2021).

We propose a novel deep neural network structure called the multispectral semantic segmentation network (MSNet) for extracting image features and spectral features in full-band data of multispectral remote sensing images. The network is a two-branch parallel structure that splits the multispectral remote sensing images into two groups of visible and invisible light, adopting ResNet-50 to simultaneously extract the image features of features in the visible band and the spectral features in the invisible bands in both encoding stages. In the decoding stage, cascaded upsampling is used to recover the feature map resolution, and the feature pyramid structure is used to fuse the multi-scale image features and spectral features layer by layer to derive the semantic segmentation results. The main contributions of this paper are:

(1) Designed a semantic segmentation network for multispectral remote sensing images: MSNet, which effectively fused image features with spectral features, and was verified by experiments in the WHU GID and the ISPRS Potsdam, two public datasets. MSNet has competitive performance;

(2) Validating the active role of spectral feature information provided by invisible light bands and spectral indices, such as NIR, DSM, NDVI, and NDWI, in the semantic segmentation task of remote sensing images;

(3) A method of dynamically synthesizing multi-channel remote sensing image datasets is proposed: the method uses the spectral indices derived from band calculation to increase the amount of spectral feature information of remote sensing image data, and the band calculation is performed simultaneously to generate spectral indices NDVI and NDWI during the loading process of the original multispectral remote sensing image dataset consisting of RGB and NIR, which is dynamically synthesized into a six-channel remote sensing image dataset. In this data pattern, MSNet has excellent segmentation accuracy, which further proves the advantages of spectral features in the semantic segmentation task of remote sensing images.

2. Related work

2.1. Semantic segmentation of remote sensing images

With the increasing spatial resolution of optical remote sensing images, the cases of “same spectral of different substances” and “different spectral of the same substances” in the images appear more frequently, and the traditional image segmentation methods based on shallow and mid-level features have limited space for improvement and are facing new challenges. With the development of artificial intelligence and deep learning, segmentation methods based on deep features have been gradually migrated to optical remote sensing images, and improved based on the characteristics of remote sensing images that are different from digital pictures, such as multispectral, highland object complexity, high inter-class similarity, and intra-class dissimilarity (Panboonyuen et al. 2019), among which the prominent representative is the segmentation model based on convolutional neural network (CNN), which realizes image segmentation by learning the features of the optical remote sensing image layer by layer while training the neural network classifier to classify the
image at the pixel level (Waldner and Diakogiannis 2020). The full convolution network (FCN) (Long, Shelhamer, and Darrell 2015) semantic segmentation model was presented in 2015, which is of epoch-making significance for image segmentation and realizes pixel-level image semantic segmentation. It replaces the full connection layer used for classification mapping in CNN structure with convolution layer, and up samples the obtained characteristic heat map to the size of the original input image through deconvolution operation, meanwhile combined with the information of the middle pooling layer, the image prediction segmentation map is generated.

Based on the FCN structure of subsampling and then upsampling, a deep convolution encoder-decoder structure SegNet (Badrinarayanan, Kendall, and Cipolla 2017) is used for target recognition and segmentation in cities such as streets and vehicles. Furthermore, UNet (Zhao et al. 2017) structure is widely used in remote sensing image segmentation tasks because of its more accurate segmentation performance in the case of less training data. DeconvNet (Trias-Sanz, Stamon, and Louchet 2008) also adopts a similar encoder-decoder idea to up-sample images to their original size, and it utilizes a deconvolution layer to densify the up-sampled of the sparse feature map instead of pooling operations. Dilated convolution is also commonly used to alleviate the contradiction between feature maps and perceptual field sizes. It uses the null convolution kernel to convolve the input image, which is able to obtain feature maps of different scales while ensuring that the perceptual field is not reduced, thus obtaining more contextual information. Typical representatives of the image segmentation structures developed by this theory are DeepLab (Chen et al. 2018) and PSPNet (Kampffmeyer, Salberg, and Jenssen 2016). In 2017, PSPNet not only applied the null convolution to ResNet but also additionally added the pyramid pooling module to better perform multi-scale context aggregation and global information acquisition to assist remote sensing images for scene parsing.

A segmentation method based on graph theory and selective search method (Guo et al. 2018) is used to enhance the training data and refine the boundaries of the segmentation results using the conditional random field (CRF) (Guo et al. 2017). Based on DeepLab V3+, adopted improved the atrous spatial pyramid pooling (ASPP), fully connected fusion paths and pre-trained encoders are used to segment the high-resolution remote sensing images (Chen et al. 2019). For the extraction of buildings in high-resolution remote sensing images, the network structure of UNet is adopted, and the channel attention mechanism and the confrontation network are integrated into the network structure (Pan et al. 2019b). A generative adversarial network (GAN) is used for feature classification of high-resolution remote sensing images, demonstrating the performance of GAN networks for semantic segmentation of remote sensing images (Benjdira et al. 2019).

The performance of most early neural networks is highly dependent on the number of training samples, which is a strongly supervised semantic segmentation method, but the labeling of training samples is a task with high time and economic costs, and the manual labeling is also subjective and uncertain. When encountering the situation that the original samples of remote sensing images are not sufficient, such as hyperspectral remote sensing data, the performance of strongly supervised semantic segmentation will be limited. Therefore, it is an effective solution to fuse the unique spectral feature information of remote sensing images and make full use of and deeply mine various types of feature information of limited samples. Meanwhile, more and more research scholars are focusing on the semantic segmentation methods for remote sensing images under weak and unsupervised supervision. Instead of using the high-cost pixel-level annotation information, the model is trained by the auxiliary annotation information of easily available boxes, lines, and points in the image or by using the semantic annotation data automatically generated by the computer, but most of the current segmentation results are not as good as the strongly supervised methods. In general, strongly supervised remote sensing image segmentation methods based on improvements are still the mainstream, and weakly supervised and unsupervised techniques still need a lot of research breakthroughs.

### 2.2. Data fusion semantic segmentation

A common approach to improving deep convolutional neural networks is to fine-tune the network pre-trained on a huge RGB scene dataset using a migration learning strategy, but this scheme is not applicable to multispectral and hyperspectral remote sensing images. A deep fully convolutional neural network (DFCNN)-based method is proposed, which uses residual correction to
fuse data from heterogeneous sensors (optical and laser) for semantic segmentation of remote sensing images; this method lays the foundation for related research (Audebert, Le Saux, and Lefèvre 2017). The idea of integrating geometry and spectral information to improve segmentation accuracy is increasingly popular, given the availability of drastically different, geo-registered data in the remote sensing field (Yuan, Shi, and Gu 2021). Early multi-source data fusion using deep learning algorithms was designed for RGB-Depth datasets (Song, Lichtenberg, and Xiao 2015). Although there are significant differences between RGB-D and multispectral remote sensing images (Lin et al. 2020), these practical methods have significantly helped in subsequent multi-spectral remote sensing image processing. An end-to-end deep convolutional network was proposed to extract the boundary information of objects in remote sensing images, and then the extracted boundary features were fed into FCN together with color, DSM (Digital Surface Model) information for network training for semantic segmentation of remote sensing images, and good segmentation results were achieved (Marmanis et al. 2018). A new deep superpixel-wise convolutional neural network called DeepQuantized-Net, which improves object patterns using superpixels instead of pixels, reduces computational cost reduction, and achieves accurate segmentation results in IND dataset (a New RGB-Depth Data Set for Building Roof Extraction) (Khoshboresh-Masouleh and Shah-hosseini 2021). FuseNet (Hazibas et al. 2017) is a new CNN network that contains two encoders and can extract features from RGB and depth images simultaneously. A new deconvolutional network is designed to process RGB and depth images separately in both encoding and decoding stages, enhancing shared information by borrowing features from each other to achieve competitive segmentation accuracy (Wang et al. 2016). MFNet (Ha et al. 2017), a convolutional neural network architecture based on the RGB-Thermal dataset, combines RGB images and thermal infrared images for semantic segmentation of self-driving street images, and the accuracy of segmentation is significantly improved by adding thermal infrared information. The RIT-18 dataset (Kemker, Salvaggio, and Kanan 2018), a dataset synthesized from real multispectral images, demonstrates that synthetic multispectral images can be used to assist in the training of an end-to-end semantic segmentation framework when there is not enough labeled image data. CoinNet (Pan et al. 2019a) can take full advantage of the initial parameters in the first convolutional layer of the pre-trained network, and comparative experiments on multispectral datasets demonstrate the effectiveness of the proposed improvements.

TU-Net and TDeepLab (Iwashita et al. 2019) combine visible and thermal images, and the implicit training network is robust to light variations. To achieve robust and accurate semantic segmentation of self-driving cars, using thermal images and fusing RGB and thermal information into a new deep neural network RTFNet (Sun, Zuo, and Liu 2019), ResNet was used for feature extraction, and a new decoder was developed to recover the feature map resolution with excellent results, and in a follow-up study they used six visual feature patterns on the existing state-of-the-art (SOTA) single-mode and data fusion semantic segmentation CNNs for evaluation, and proposed a dynamic fusion module (DFM) that can be easily deployed in existing data fusion networks to effectively and efficiently fuse different types of visual features, and the resulting DFM-RTFNet (Wang et al. 2021) achieved competitive performance in road benchmark tests.

At present, the research on semantic segmentation based on data fusion is still mainly in the field of digital pictures, and the “Y” type network structure of simultaneous coding of RGB images and thermal infrared images with the merging and decoding of feature maps has been established to achieve the significant improvement of data fusion and segmentation accuracy at the feature level. In the research of the remote sensing image field, multi-source remote sensing data fusion at the pixel level is mainly based on optical remote sensing image and DSM, LiDAR, and SAR. However, mining the spectral feature information of remote sensing image and fusing the visible light bands and invisible light bands, and even spectral index information at the feature level also has research value and potential.

3. Methods

The multispectral semantic segmentation network for remote sensing images designed in this paper aims to utilize all the band information of remote sensing images. Through band splitting and recombination, independent downsampling of visible light and invisible light bands is performed, and the image features of objects in the visible
light band and the spectral features of the invisible light band are simultaneously extracted and merged. The structure of this dual-branch synchronous encoding is to receive different types of data simultaneously, and to directly invoke the pre-trained model for deep-level feature extraction. In the decoding process, the feature map is fused layer by layer to achieve high-precision semantic segmentation results of remote sensing images.

3.1. The overall architecture

We present a new deep neural network structure called MSNet for multispectral remote sensing images for semantic segmentation of multiple feature scenes. Figure 1 shows the overall structure of MSNet, which mainly consists of two parts: 1) Band splitting and simultaneous feature extraction, split the multispectral remote sensing image bands into visible and invisible light, and both draw support from ResNet-50 for feature extraction in the coding stage; 2) Feature fusion decoding, adopt the cascaded upsampling method to recover the feature map resolution in the decoding stage, and uses the feature pyramid structure to layer-by-layer fusion of multi-scale image features and spectral features in the upsampling process to finally obtain semantic segmentation results.

3.2. Bands splitting and simultaneous feature extraction

We split the bands of multispectral remote sensing images into two groups, visible and invisible light, and then carry out grouped synchronous feature extraction, where the visible band part is the RGB of conventional digital pictures, mainly used for extracting color features, texture features, shape features, and spatial relationship features; the invisible light band part contains two data types, one is the NIR band of the original multispectral remote sensing images, and the other is the data. The spectral indices NDVI, NDWI, etc., which are synthesized synchronously during the loading process, are used to extract the spectral features, so MSNet has two types of multispectral data pattern (4-channel, 4c) and dynamically synthesized multichannel data patterns (6-channel, 6c), and Figure 1 shows MSNet in 6c data pattern. This dual-branch synchronous coding structure is the current conventional scheme of multi-source data fusion semantic segmentation model. The purpose is to receive different types of data simultaneously and can directly call the pre-trained model to extract deep-level features. Simultaneously, it also draws on the fundamental hypothesis behind the inception that cross-channel correlations and spatial correlations are sufficiently decoupled that it is preferable not to map them jointly (Chollet 2017).

![Figure 1. The overall architecture of MSNet, dynamic synthesis of multichannel data patterns (6-channel, 6c).](image-url)
The ResNet-50 is used in the group synchronization coding stage for feature extraction; while taking into account the deep feature extraction and model training efficiency, it is worth noting that the number of channels should be initialized according to the number of bands of the data pattern; the remote sensing images have a lower resolution compared with the digital pictures, and the pooling layer causes information loss and ignores the correlation between the whole and the parts, leading to further reduction in resolution, some models in the design. The pooling layer is removed in the design of some models to solve the above problems, but simultaneously it significantly increases the computational cost of the model.

3.3. Image feature and spectral feature fusion decoding

We present MSNet for simultaneous encoding and decoding of multispectral remote sensing images by band grouping similar to the method of the siamese neural network (Iwashita et al. 2019), but they usually only merge the results of the last stage of the sub-network decoder. We have borrowed from other excellent network processing methods, U²-Net uses the saliency fusion module to generate saliency probability maps at the side of each decoder level (Qin et al. 2020) and UPerNet fuses multiple semantic level features using feature pyramids (Xiao et al. 2018). Figure 2 shows the fusion decoding process, which retains the results of each layer of the subgroup decoder and uses a feature pyramid structure to fuse the multi-scale image features and spectral features from the upsampling process layer by layer for feature fusion decoding, and finally obtains the semantic segmentation results.

The left and right decoders represent the upsampling process of RGB and NIR, NDVI, NDWI synthesized bands, respectively, and they adopt a jump connection using channel splicing, which allows the combination of deep, semantic, and coarse feature maps in the decoder and shallow, low-level, and fine feature maps in the encoder. The middle part is a feature pyramid structure, which also uses channel splicing to fuse the multi-level feature results of the band

Figure 2. Image feature and spectral feature fusion decoding.
grouping in both decoders step by step and keeps the upsampling synchronized. This approach is effective in remote sensing image resolution tasks that take into account scene, texture, and material, but also annotation blending and lack of significant depth-of-field contrast. We use the pixel shuffle method (PixelShuffle) (Shi et al. 2016) in the decoding stage of the model, which was originally proposed to solve the image super-resolution problem by taking the low-resolution feature maps and up-sampling them to obtain high-resolution images by convolution and multi-channel inter-reconfiguration.

4. Experiments

4.1. Multispectral remote sensing image datasets

In order to verify that msnnet proposed by us has competitive performance and model generalization ability in the semantic segmentation task of multispectral remote sensing images, two public multispectral remote sensing image datasets WHU GID (Tong et al. 2020) and ISPRS Potsdam (“2D Semantic Labeling –Potsdam,” n.d.) are selected, and three images are selected, respectively, in the two datasets as test data. Table 1 shows the details of data sets and preprocessing.

Geospatial data abstraction library (GDAL) (“GDAL –GDAL documentation,” n.d.) is used to load multispectral remote sensing image data. According to the requirements of different network models for the number of data channels, the multispectral remote sensing image is divided into bands, band sequence is adjusted, spectral index calculation and multi-channel data synthesis are performed. And other operations to achieve the function of dynamic loading of remote sensing image datasets. The above data processing is realized in memory, so the preprocessing data only require image slicing, and no additional dataset organization and arrangement is needed for the specific network model experimental requirements, which avoids a large amount of redundant data occupying storage space and also improves the utilization efficiency of remote sensing datasets. We divide the training validation process of the network model with the premise that the datasets are grouped in different loading modes as follows:

1. 3c data pattern: applicable to digital pictures semantic segmentation networks such as UNet, loading complete 4-band multispectral remote sensing images, but only reading visible light RGB 3-band.

2. 4c data pattern: applicable to visible and invisible light fused semantic segmentation networks such as RTFNet, loading and reading complete 4-band multispectral remote sensing images, with the bands split into RGB + NIR. It should be noted that the original band order of the GID dataset is NIR-R-G-B, and the band order is adjusted to R-G-B-NIR in order to maintain the consistency of the model calculation.

3. 6c data pattern: In order to dig deeper into the spectral feature information of remote sensing images, when loading and reading 4-band multispectral remote sensing images, band operations are performed to generate spectral indices: normalized difference vegetation index (NDVI) (Rouse et al. 1974) and normalized difference water index (NDWI) (Gao 1996), which are synthesized into the 6-channel dataset.

NDVI and NDWI were selected for data synthesis because regardless of the two publicly available datasets used in this paper or other optical remote sensing images, these two spectral indices have a wide range of applicability and can be obtained directly by band calculations, and the combined reflectance operations of features in multiple wavelength ranges enhance the characteristics of vegetation and water.

It should be noted that there is no water in the Potsdam dataset, so NDWI will not be calculated, but the dataset contains DSM, which can be used as

| Table 1. Datasets and data preprocessing details. |
|-----|-----|-----|-----|-----|-----|
| **Datasets** | **Image resolution** | **Number of bands** | **Band sequence** | **Image slice size** | **Number of image slices** | **Label designation** |
| GID | 1 m | 4 | NIR-R-G-B | 512 × 512 | train:2992 | clutter, water, farmland, meadow, forest, built-up |
| | | | | | val:1088 | low_veg, tree, imp_surf, building, car, clutter |
| | | | | | train:4225 | |
| | | | | | val:1352 | |
a spectral data supplement. DSM is a ground elevation model that includes the height of surface buildings, bridges, trees, etc., based on digital elevation mode (DEM), which further covers other surface information besides the ground, and is the most realistic representation of the ground undulation. After dynamic synthesis of the 6-channel dataset, it is then split into two groups: the GID dataset is RGB + NIR-NDVI-NDWI, and the Potsdam dataset is RGB + NIR-NDVI-DSM.

4.2. Implementation details

We designed MSNet based on the PyTorch 1.8.0 framework, and the network model training was performed on a GPU server: containing 2 CPUs (Intel Xeon E5-2640 v4) with a total of 128 GB RAM, 4 GPUs (NVIDIA Tesla V100 16 GB) with a total of 64 GB VRAM. The main parameter settings: the data batch size is set according to different networks to ensure maximum memory utilization, the number of threads is 40, the initial learning rate: 1e-4, the dynamic adjustment strategy of learning rate: ReduceLROnPlateau, the optimizer: AdamW (Loshchilov and Hutter 2018), the cross-entropy loss function is selected for the multi-category labeled dataset, and the number of training epochs: 100. The normalization parameters in the experiments of this paper: the mean and standard deviation of each channel, instead of using the default parameter values in the semantic segmentation of digital pictures, are pre-calculated based on the corresponding multi-spectral remote sensing image datasets.

4.3. Evaluation metrics

The evaluation work is divided into two parts: segmentation accuracy evaluation and network model efficiency evaluation.

To enable an objective and fair cross-sectional comparison with other proposed network models, six common evaluation criteria are used, including Overall Accuracy (OA), Precision, Recall (Davis and Goadrich 2006), F1-Score (F1) (Huang et al. 2015), Intersection over Union (IOU) and Frequency Weighted Intersection over Union (FWIoU). The results of these evaluation criteria were calculated using the confusion matrix, where TP represents the number of True Positive classes, TN represents the number of True Negative classes, FP represents the number of False Positive classes, and FN represents the number of False Negative classes:

\[
\text{OA} = \frac{TP + TN}{TP + TN + FP + FN}
\]

Precision measures the proportion of positive samples detected by the model that are correct, and Mean Precision (MP) represents the average accuracy across all categories:

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

Recall measures the proportion of correct predictions by the model in all positive samples.

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

F1-Score (F1) is the combined consideration of Precision and Recall:

\[
F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FN + FP}
\]

Intersection over Union (IoU) measures the ratio of the intersection and union between the predicted result of a certain category and the true value of the model. Mean Intersection over Union (mIoU) represents the ratio of the intersection and union between the predicted result of each category and the true value of the model:

\[
\text{IoU} = \frac{TP}{TP + FP + FN}
\]

Frequency Weighted Intersection over Union (FWIoU) sets the weights according to the frequency of each category, and multiplies the IoU of each category with the weights to sum:

\[
\text{FWIoU} = \frac{TP + FN}{TP + TN + FP + FN} \times \frac{TP}{TP + FP + FN}
\]

The inferential relationship between evaluation indicators:

\[
\frac{1}{\text{Recall}} = 1 + \frac{1}{\text{IoU}} - \frac{1}{\text{Precision}}
\]
The network model efficiency evaluation mainly includes the overall model training time, model save size, maximum batch size with a total of 64GB VRAM, number of model parameters, floating point operation speed, inference time, and inference speed.

4.4. Comparative with the state-of-the-art methods

We use the GID and Potsdam datasets, respectively, to experimentally compare MSNet of both data models with some mature semantic segmentation network models DeepLabV3+, FPN, PSPNet, UNet, and RTFNet. All models uniformly use ResNet-50 as the backbone network, which is used to improve the efficiency of feature extraction while ensuring the fairness of the experimental comparison. Considering that RTFNet has been verified to have significant advantages over ERFNet (Romera et al. 2018), PSPNet (Zhao et al. 2017), SegNet (Badrinarayanan, Kendall, and Cipolla 2017), DUC-HDC (Wang et al. 2018), MFNet (Ha et al. 2017) and FuseNet (Hazirbas et al. 2017) equivalent networks in the task of processing 4-band images, no repeatability validation experiments are conducted in this paper.

4.4.1 Comparison of training stages

Figure 3 and Figure 4 respectively show the validation results and trend of the overall evaluation index value of each network model during the training process in the GID and Potsdam datasets. Table 2 and Table 3 respectively represent the segmentation accuracy evaluation of each network model for all label categories after the final round of training, and the overall validation results.

In the validation evaluation results of the training phase of each network model, except for the segmentation index of the meadow change greatly, the variation of the segmentation accuracy and intersection ratio of other categories in each network model is within 6%, which is caused by the uneven number and scale of feature categories contained in each remote sensing image in the GID dataset, and the proportion of images containing labels of all feature categories. The proportion of images containing all feature category labels is less than 30%, and the sample size of meadow is insufficient. Combined with the results of the overall evaluation index values, MSNet (6c) has the best overall performance.

In the validation evaluation results of the training phase of each network model in the Potsdam dataset, all categories were better segmented and identified without significant differences. In the overall evaluation and the detail evaluation results for distinguishing the category labels, the results showed a trend of DeepLabV3+(3c) < PSPNet(3c) < FPN(3c) < UNet(3c) < RTFNet(4c) < MSNet(4c) < MSNet(6c).

The validation results from two datasets, Tables 2 and 3 show that our proposed MSNet achieves the best results in both sub-indexes and overall indexes in 4c and 6c data patterns, especially the highest accuracy in 6c data pattern. These results prove the performance of MSNet and also validate the effectiveness of our proposed method for the dynamic synthesis of multichannel remote sensing image datasets.

Figure 3. Comparison of the validation results of the network models in the training phase in the GID dataset.
Table 2. Details of the validation results of the network models in the GID dataset.

| GID      | Precision | IoU   | Precision | IoU   | Precision | IoU   | Precision | IoU   | Recall | F1  | mIoU  | FWIoU |
|----------|-----------|-------|-----------|-------|-----------|-------|-----------|-------|--------|-----|-------|-------|
| DeepLabV3 | 79.15     | 64.04 | 118.6     | 70.86 | 29.23     | 19.77 | 70.62     | 58.50 | 77.93  | 64.45| 77.09 | 68.37 |
| +6c)     | 91.18     | 94.36 | 90.66     | 83.27 | 118.6     | 70.86 | 29.23     | 19.77 | 70.62  | 58.50| 77.93 | 64.45 |

Table 3. Details of the validation results of the network models in the Potsdam dataset.

| Potsdam | low_veg | tree | imp_surf | building | car | clutter |
|---------|---------|------|----------|----------|-----|---------|
| Precision | IoU   |       | Precision | IoU   |       | Precision | IoU   |       | Precision | IoU   |       | Precision | IoU   |       | Precision | IoU   |       | Precision | IoU   |       | Precision | IoU   |       |
| DeepLabV3 | 80.34 | 63.32 | 61.23     | 48.44   | 85.31 | 72.25 | 68.96     | 62.61 | 49.45  | 38.63| 79.84 | 76.07 |
| +3c)     | 92.68 | 85.67 | 90.49     | 83.54   | 94.75 | 88.93 | 97.10     | 94.49 | 90.61  | 82.93| 73.65 | 93.51 |

The 4-channel data composed by adding NIR bands have increased the spectral characteristics and expanded the amount of remote sensing data information compared with RGB in visible bands, and the segmentation accuracy was significantly improved in the validation results of both RTFNet and MSNet (4c). The 6-channel data synthesized by band calculation add spectral indices NDVI and NDWI, which are more consistent with the conventional data processing and feature segmentation methods of remote sensing images. The segmentation accuracy of MSNet (6c) is further improved, which proves the positive role of spectral feature information in the semantic segmentation of remote sensing images. Figure 3 and Figure 4 show the specific validation results of the training process of each network model in the GID and Potsdam datasets, respectively, where the evaluation criteria curves of MSNet show a faster upward trend and stable fluctuations.

4.4.2. Comparison of test stages

In order to further compare the actual segmentation accuracy of the model, we randomly selected three images from each of the two datasets for testing. Generated the comparison diagram between the segmentation results of each model and the truth label, and established the confusion matrix to quantitatively evaluate the segmentation accuracy of each category. MSNet is already in the leading position in the 4c data pattern, and in the 6c data pattern, most of the index values have further improved slightly except for the IoU of meadow and the mIoU of the overall index, especially in the segmentation accuracy of vegetation and artificial objects. Figure 5 shows the visualization status and ground true value of the test data in all channels of the GID dataset. The segmentation of farmland is more accurate and more clearly separated from the road. In terms of water segmentation, bridges over rivers are also effectively distinguished. The above typical results demonstrate that the 6-channel data synthesized by adding NDVI and NDWI derived from band calculation can achieve excellent segmentation results in MSNet.

Figure 6 shows the quantitative evaluation of the prediction result atlas, which is generated from the GID test dataset in each network model (Figure 3), and the results of the overall quantitative evaluation of each metric in the test phase are consistent with the evaluation results in the training validation phase (Table 2), which further confirms the competitive performance of our proposed MSNet.

Figure 6 includes the detailed evaluation comparison results of the categorical labels. It is worth stating that the reason for the absence of meadow and forest data in the first set of statistical plots is the absence of these two categories in the randomly selected test image.
The evaluation result of Precision and IoU for each category obtained from the predictions of each network model in the GiD test dataset can be compared with the table of evaluation results in the training validation stage (Table 2), and some different phenomena can be found:

1. The results of IoU values for each category in the two stages of training validation and testing are consistent and coincide with the overall evaluation results;
2. The results of Precision values for each category in the training phase, are some differences between the results of training validation and testing; specifically, the comparison trend of the precision rate results of each category in the training validation stage is UNet(3c) < RTFNet(4c) < MSNet(4c) < MSNet(6c); however, in the testing stage, UNet(3c) has the highest prediction precision rate for clutter, and RTFNet(4c) has the highest prediction precision rate for water, farmland, and meadow, and MSNet(4c) and MSNet(6c) had the highest prediction precision for the forest;
3. Comparing the Precision and IoU values of each category, some results of the same category predicted by the same network model showed high Precision values and low IoU values.

The reasons for this phenomenon mainly include the problem of evaluation indicators: in this paper, six indicators are selected in the overall evaluation, and two indicators are selected in the categorical evaluation, in order to prevent the limitation of single indicator evaluation. Precision focuses on evaluating the exactness of the prediction results, indicating the proportion of correct predictions, and is generally used in conjunction with Recall, which focuses on evaluating the completeness of the prediction results, indicating the proportion of all predictions, but in this paper, we selected a more intuitive evaluation index IoU instead of Recall, and the inference relationship between these two indicators is detailed in Equation 8. The phenomenon of high Precision value and low IoU value indicates that the model predicts the category with a high precision rate but is incomplete. From Figure 6, we can find that the prediction precision of UNet(3c) for water in the first two images is less than 20%, which is significantly lower than other network models, and the prediction of RTFNet(4c) for forest is also lower than the average of all network models. MSNet of both models has no outstanding prediction performance in terms of Precision value, but IoU value performs well and has a more complete ability to predict the extracted categories, which is also confirmed by Figure 5.

The problem of samples: Same as the problems of the number and proportion of categories of samples mentioned in the training validation stage, the GiD dataset consists of GF-2 remote sensing images of more than 60 different cities in China, and the category labels are not finely outlined and labeled, but synthesized with land cover data, so this dataset has the phenomenon of completely different feature and landform characteristics within the same category, also including the unbalanced proportion of samples within each category label. The imbalance in the proportion of samples within each category of labels causes the situation that most images
contain only some of the feature categories. In the training and validation phase of the network model, a larger number of images were used and iterated in the network model continuously in the form of sliced samples, which largely weakened the above problems, but in the testing phase, there were only three randomly selected images, and when the sample size was small, the problems existing in the dataset will be highlighted.

In the Potsdam dataset validation results, MSNet also achieved the best results with further improvement of all metric values in the 6c data pattern, and since there is no water in the Potsdam dataset, but there are DSM data, DSM was used to replace NDWI in the 6c data synthesis. Figure 7 shows the visualization status and ground true value of the test data in all channels of the Potsdam dataset, as well as their prediction result plots in each network model. The segmentation effect of vegetation is improved more obviously, while the addition of DSM data drives the accuracy of house and road identification.

Figure 8 shows the quantitative evaluation of the prediction result atlas, which is generated from the Potsdam test dataset in each network model (Figure 4), and the results of the overall quantitative evaluation of each metric in the test phase are consistent with those in the training validation phase (Table 3), further confirming that MSNet has better performance.
Figure 8 includes the detailed evaluation of comparison results of the categorical labels. In the Precision value, which reflects the exactness of the prediction results, the prediction levels of each network model are similar for different test image categories, and the prediction precision of MSNet(6c) is slightly higher for imp_surf; In the IoU value section that characterizes the completeness of the predicted results, MSNet(6c), which incorporates the spectral features of NDVI and DSM, has a clear advantage. The best results were achieved in all categories of completeness prediction; meanwhile, the prediction evaluation indexes the Precision and IoU value of each category in different network models were coordinated and stable, without the opposite situation similar to that in the rating results of GID test dataset, mainly because the image acquisition of Potsdam dataset was concentrated in one city, and the features and landforms within the same category did not change significantly. The relatively balanced proportion of samples within each category label also confirms to some extent that the GID dataset has problems such as labeling errors.

The validation results of the two datasets can be summarized: 1) Our proposed MSNet for semantic segmentation of multispectral remote sensing images has competitive performance; 2) In the NIR band, water has higher spectral absorptivity values and vegetation and artificial features have higher spectral reflectance values, these spectral features facilitate the task of semantic segmentation of remotely sensed images; 3) Passed the spectral index NDVI and NDWI obtained by band calculation are combined with 6-channel data, which helps to improve the accuracy of semantic segmentation of multispectral remote sensing images, especially the identification of vegetation, water, and artificial objects.

4.5. Ablation study

In the ablation study, we mainly verify the effects of data group encoding and decoding methods on segmentation results, and the main results when using the GID dataset are shown in Table 4:

1) Encoding method, the encoding method of splitting the bands into two groups in 4c and 6c data patterns can reduce the training time and improve the training efficiency, and increasing the number of filters has less effect on the segmentation results;
2) Decoding structure, the conventional fusion decoding method is to sum up the feature maps of branching structure directly, MSNet adopts feature fusion in the decoding part,
that is, two groups of encoding results representing image features and spectral features are decoded independently, and then the feature pyramid structure is used to fuse two groups of independent decoding layer by layer, Table 4 shows that the final results of decoding by feature fusion are generally better;

(3) Spectral index information, the ablation experimental results also confirmed the influence of spectral index information such as NDVI and NDWI derived from band calculation on the semantic segmentation results, and the spectral index information in MSNet_2 with the conventional fusion decoding method and MSNet (6c) with the feature fusion decoding method, resulting in the segmentation accuracy showing completely opposite results.

In addition, we tried different band combination methods in the 4c data pattern in the ablation experiment: replacing NIR with NDVI or NDWI in the GID dataset and replacing NIR with DSM in the Potsdam dataset, and the highest accuracy of each band combination method in the final model validation results was RGB + NIR, the possible reason is that the NDVI and NDWI directly derived from the band calculation are not post-processed such as threshold splitting, and there are some incorrect spectral index results. The DSM in the
4.6. Evaluation in efficiency

Model efficiency evaluation has received extensive attention in deep learning research, with emphasis on the inference speed of the network model in the prediction phase, however elements such as the overall time consumption and computational cost of the network model during the training phase also deserve attention. Hardware devices, especially GPUs, are necessary for deep learning research. For this reason, GPU servers consisting of multiple GPUs for parallel computing, a larger number of CUDA cores and video memory capacity have been very commonly used in the field of deep learning. The complexity of the model determines the number of samples loaded per batch and the computation time, we also tried to use more complex backbone networks for deeper mining features in our study, such as ResNet-101 and ResNet-152, using the GID dataset and comparing the results of the model training validation phase, the improvement of the evaluation metrics is not obvious (Table 5), mainly due to the lack of sample labeling accuracy and unbalanced distribution of sample categories. However, the consequent training time increased significantly, leading to higher computational cost and lower computational efficiency. The result is a significant increase in training time,

Table 4. Results of ablation study.

| Methods                      | batch_size (max) | runtime | OA     | MP     | Recall | F1     | mIoU  | FWIoU |
|------------------------------|------------------|---------|--------|--------|--------|--------|-------|-------|
| MSNet_1 4c, RGB + NIR,       | 36               | 2 h 57 m| 93.39  | 90.43  | 92.34  | 91.38  | 84.12 | 87.66 |
| 2 groups of encoders,        |                  |         |        |        |        |        |       |       |
| Conventional decoder        |                  |         |        |        |        |        |       |       |
| MSNet_2 6c, RGB + NIR-NDVI-NDWI | 36            | 3 h 34 m| 93.37  | 90.80  | 91.13  | 90.97  | 83.43 | 87.63 |
| 2 groups of encoders,        |                  |         |        |        |        |        |       |       |
| conventional decoder        |                  |         |        |        |        |        |       |       |
| MSNet_3 6c, RGB + NIR + NDVI + NDWI | 20          | 5 h 16 m| 92.78  | 89.76  | 90.44  | 90.10  | 81.98 | 86.61 |
| 4 groups of encoders,        |                  |         |        |        |        |        |       |       |
| Conventional decoder        |                  |         |        |        |        |        |       |       |
| MSNet_4 4c, RGB + NIR,       | 32               | 3 h 30 m| 92.40  | 89.63  | 90.18  | 89.90  | 81.66 | 85.97 |
| 2 groups of encoders,        |                  |         |        |        |        |        |       |       |
| Double filters,              |                  |         |        |        |        |        |       |       |
| Conventional decoder        |                  |         |        |        |        |        |       |       |
| MSNet(4c) 4c, RGB + NIR,     | 40               | 2 h 42 m| 93.49  | 90.13  | 92.92  | 91.50  | 84.34 | 87.81 |
| 2 groups of encoders,        |                  |         |        |        |        |        |       |       |
| Feature fusion decoder       |                  |         |        |        |        |        |       |       |
| MSNet(6c) 6c, RGB + NIR-NDVI-NDWI | 40          | 3 h 5 m | 93.85  | 91.39  | 91.49  | 91.44  | 84.23 | 88.48 |
| 2 groups of encoders,        |                  |         |        |        |        |        |       |       |
| Feature fusion decoder       |                  |         |        |        |        |        |       |       |

Potsdam dataset can be applied to directly identify houses and trees with significant height information, but the height differences between other low features are not obvious and difficult to distinguish.
resulting in higher computational cost and lower computational efficiency. It is not worth advocating the approach of blindly designing overly complex models and exchanging high computational cost for small accuracy improvement.

The results of Table 6 show a comparison with UNet (3c) and RTFNet (4c) under the same hardware environment and operating parameters while ensuring the highest video memory utilization. Model training and validation time (runtime), model size, maximum batch size, model parameters, floating point operations (FLOPs), and frames per second (FPS) are selected as the measurement metrics. The model efficiency of our proposed MSNet is at the average level in both 4c and 6c data patterns, and when combined with the model testing results, it can be confirmed that MSNet has favorable segmentation accuracy, while the computational cost and inference speed are balanced.

5. Discussion

In the testing phase, we selected three sets of images and truth labels from the two datasets to visually compare the predicted segmentation result graphs of each network model (Figure 5 and Figure 6). The confusion matrix is again used to calculate the values of each evaluation metric in the predicted segmentation result maps to quantitatively evaluate the test results, especially the degree of segmentation in each category, which is an effective way to objectively reflect the true performance of the network models.

5.1. Dual-branch model structure

In the field of natural images, pre-training models such as ResNet help to accelerate neural network training and improve the training results simultaneously, mainly because the parameters in pre-training models are trained from massive data such as ImageNet and have generic features, and most of the migration learning based on this has been effective (He, Girshick, and Dollar 2019). However, these parameters are based on RGB three channels, which are not applicable to the processing of remote sensing images with multiple wavebands, and we also conducted comparative experiments for this purpose, Table 7 shows the test results of three images in the GID dataset, without changing the original structure of UNet, and increasing the number of channels to load NIR, NDVI, and NDWI wavebands to form the UNet(4c) and UNet(6c) networks, but the indicators in their test results are significantly lower than those of UNet(3c), which only uses the RGB band. When the input data have different structures, separate networks are usually used to process each data type and fuse them in the classification stage; therefore, dual-branch model structure has been used in past multi-source heterogeneous data fusion (Yuan, Shi, and Gu 2021), to maintain the advantages of the pre-trained model for RGB bands while using other data as auxiliary features to improve the model recognition accuracy.

The multi-source data fusion model, represented by RTFNet and others, uses dual-branch structure for grouping and extracting features in the encoding stage. The

| Table 5. Experimental results for different backbone network configurations. |
|-------------------------|-----------------|---------|---------|---------|---------|---------|---------|---------|
| Backbone Methods | batch_size (max) | runtime | OA   | MP    | Recall | F1   | mIoU   | FWIoU  |
|-------------------|------------------|---------|------|-------|--------|------|--------|--------|
| MSNet(6c)         | ResNet-30        | 40      | 3 h 5 m | 93.85 | 91.39  | 91.49 | 91.44  | 84.23  | 88.48  |
| MSNet(6c)         | ResNet-101       | 26      | 5 h 48 m | 94.38 | 91.77  | 91.77 | 91.77  | 84.79  | 88.63  |
| MSNet(6c)         | ResNet-152       | 16      | 8 h 43 m | 94.49 | 92.24  | 91.69 | 92.01  | 85.21  | 88.81  |

| Table 6. Efficiency evaluation results. |
|------------------------|-----------------|---------|---------|---------|---------|---------|---------|
|                         | GID         | Potsdam | model_size | batch_size (max) | Parameters | FLOPs  | FPS    |
| DeepLabV3+(3c)       | 3 h 20 m    | 4 h 25 m  | 209 MB    | 36       | 54.7     | 83.11  | 31.46  |
| FPN(3c)              | 2 h 26 m    | 2 h 28 m  | 274 MB    | 64       | 27.15    | 62.32  | 35.11  |
| PSPNet(3c)           | 1 h 49 m    | 2 h 24 m  | 196 MB    | 48       | 46.58    | 26.36  | 44.15  |
| UNet(3c)             | 2 h 16 m    | 3 h 3 m   | 150 MB    | 48       | 37.34    | 83.73  | 29.14  |
| RTFNet(4c)           | 3 h 58 m    | 5 h 28 m  | 707 MB    | 24       | 185.24   | 208.98 | 12.70  |
| MSNet(4c)            | 2 h 42 m    | 3 h 33 m  | 370 MB    | 40       | 92.98    | 68.92  | 23.13  |
| MSNet(6c)            | 3 h 5 m     | 4 h 12 m  | 371 MB    | 40       | 92.98    | 69.33  | 22.98  |


network we designed extends the dual-branch structure all the way to the decoding stage, and the multi-source data are encoded and decoded simultaneously after grouping, which is more similar to the siamese neural network structure, but the data fusion part adopts a feature pyramid structure to fuse the multi-scale image features and spectral features in the upsampling process layer by layer. The results shown jointly in Tables 6, 7 and Figure 5 demonstrate the MSNet validity. Segmentation accuracy and inference speed for various feature classes differ significantly in the semantic segmentation networks selected for different combinations of data patterns, and this difference can provide empirical advice in scenarios oriented to the interpretation of specified features.

3c data pattern: The main approach in the field of computer vision, four classical models are selected in this paper, with simple network structure, fast training and inference, and better performance in the recognition of artificial features such as buildings and roads, but the phenomenon of mis-segmentation and omission is more obvious in complex ground coverage areas, for example, in the last image, none of the four models can completely segment the bridge over the water.

4c data pattern: with the addition of NIR bands and a two-branch network structure, the complexity of the model increases significantly, and this approach improves the recognition probability of vegetation and water bodies that behave sensitively in the NIR bands, and the results of RTFNet(4c) and MSNet(4c) show a significant improvement in the edge segmentation accuracy of vegetation and water.

6c data pattern: During the loading process of multispectral remote sensing images of four bands composed of RGB and NIR, simultaneous band calculation is performed to generate spectral indices NDVI and NDWI, which are dynamically synthesized into 6c data. MSNet(6c) adds a spectral index classification method to form a strengthened supervised semantic segmentation, which mainly improves the recognition accuracy of features with fragmented and fine distribution and further improves the overall accuracy.

5.2. The effect of spectral features

Previous studies have confirmed that the results of using RGB and thermal infrared band fusion in digital pictures semantic segmentation tasks are better than RGB 3-band images, while the task scenes contain classification targets with obvious temperature contrasts (Ha et al. 2017; Hazirbas et al. 2017; Wang et al. 2016). The traditional processing of remote sensing images reorganizes the NIR band with the visible band to synthesize standard pseudo-color images for visual interpretation. Based on the significant differences in the spectral reflectance of different features such as vegetation and water in the NIR band, we fuse RGB and NIR bands, which is essentially a fusion of image features with spectral features, and design a semantic segmentation network model.

Remote sensing images are converted into apparent reflectance with actual physical meaning through radiometric calibration in each band of the full spectrum. Analysis based on the full-spectrum characteristics of ground objects is the main method in the field of remote sensing data processing. The reflectance values of vegetation in the red and near-infrared bands and the absorptivity values of water in the green and near-infrared bands all have a very significant upward trend. NDVI and NDWI can eliminate most of the irradiance changes related to radiometric calibration, solar azimuth angle (SOA), terrain, cloud shadows, and atmospheric conditions, enhance the response to vegetation and water, and achieve rapid differentiation from other ground objects.

The Potsdam dataset contains the digital surface model (DSM), which provides depth-of-field information similar to that in digital pictures. The experiments in this paper only verified the effectiveness of the proposed MSNet for such special spectral information, and no more in-depth research was conducted for DSM, for instance, DSM can convert two-dimensional remote sensing images into three-dimensional images, unlike the form of data with high spectral dimension, which is a data cube with real height difference, and using 3D convolutional neural network can extract richer features of features in layers, which will help to improve the extraction accuracy of local details (Konapala, Kumar, and Khaliq Ahmed 2021; Li, Wang, and Li 2021; Roy et al. 2020).

Our experiments validate the effectiveness and advantages of semantic segmentation using multispectral remote sensing data after band combination calculation, if multispectral remote sensing images with more number of bands, for example, containing short-wave infrared and mid-wave infrared, short-wave infrared (SWIR), medium-wave infrared (MIR),
| Model                     | GF2_PMS1__L1A00010644S4-MSS1.tif | GF2_PMS1__L1A00016808S1-MSS1.tif | GF2_PMS2__L1A00014714S6-MSS2.tif |
|---------------------------|----------------------------------|----------------------------------|----------------------------------|
|                           | Precision | IoU | Precision | IoU | Precision | IoU | Precision | IoU | OA | Precision | IoU | OA | Precision | IoU | OA | Precision | IoU | OA | Precision | IoU | OA | Precision | IoU | OA | Precision | IoU | OA | Precision | IoU | OA | Precision | IoU | OA | Precision | IoU | OA | Precision | IoU | OA | Precision | IoU | OA |
| UNet(3c)                  | 79.77      | 32.28 | 17.32     | 17.25 | 90.05      | 86.50 | -             | -   | -             | -   | 75.59      | 67.47 | 74.62  | 65.68      | 69.29 | 67.44 | 50.88      | 64.74 |
| UNet(4c)                  | 75.90      | 24.44 | 4.25      | 4.25  | 79.15      | 75.24 | -             | -   | -             | -   | 68.24      | 61.56 | 64.71  | 56.89      | 44.18 | 49.74 | 33.10      | 54.18 |
| UNet(6c)                  | 82.29      | 27.54 | 19.11     | 19.08 | 79.40      | 76.36 | -             | -   | -             | -   | 60.50      | 57.55 | 67.68  | 60.33      | 37.51 | 46.26 | 30.09      | 57.50 |
| RTNet(4c)                 | 59.63      | 31.29 | 38.34     | 37.75 | 96.44      | 88.85 | -             | -   | -             | -   | 81.25      | 68.23 | 79.55  | 56.81      | 62.28 | 45.22 | 69.60      | 39.76 |
| MSNet(4c)                 | 75.93      | 34.63 | 27.61     | 27.59 | 94.42      | 89.79 | -             | -   | -             | -   | 75.36      | 68.51 | 78.38  | 68.33      | 55.44 | 61.21 | 44.10      | 68.91 |
| MSNet(6c)                 | 76.45      | 37.38 | 36.82     | 36.78 | 94.71      | 90.20 | -             | -   | -             | -   | 80.14      | 72.94 | 80.66  | 72.03      | 77.09 | 74.47 | 59.33      | 71.56 |
| UNet(3c)                  | 86.91      | 43.89 | 17.39     | 17.31 | 89.23      | 85.23 | 33.40         | 32.97 | 84.52         | 83.04 | 74.94      | 66.95 | 77.87  | 64.40      | 78.82 | 70.88 | 54.90      | 67.12 |
| UNet(4c)                  | 83.94      | 34.59 | 4.35      | 4.34  | 80.73      | 75.25 | 1.07          | 1.07  | 73.37          | 72.33 | 67.65      | 61.03 | 68.26  | 51.85      | 67.35 | 58.59 | 41.44      | 55.86 |
| UNet(6c)                  | 86.15      | 37.48 | 19.20     | 18.87 | 81.24      | 76.02 | 10.40         | 10.31 | 76.21          | 74.87 | 59.98      | 56.86 | 70.96  | 55.53      | 72.16 | 62.76 | 45.73      | 58.86 |
| RTNet(4c)                 | 77.36      | 42.56 | 38.41     | 37.78 | 94.79      | 87.27 | 41.17         | 37.28 | 75.62          | 74.91 | 80.55      | 67.73 | 78.76  | 67.98      | 79.65 | 73.35 | 57.92      | 67.93 |
| MSNet(4c)                 | 85.29      | 45.74 | 27.64     | 27.61 | 93.24      | 88.29 | 23.70         | 23.32 | 86.39          | 84.88 | 74.71      | 67.94 | 79.92  | 65.16      | 80.61 | 72.07 | 56.33      | 69.54 |
| MSNet(6c)                 | 85.34      | 47.55 | 36.87     | 36.83 | 93.50      | 88.56 | 29.75         | 29.47 | 86.08          | 84.62 | 79.45      | 72.33 | 81.31  | 80.50      | 82.66 | 74.92 | 59.89      | 71.26 |
| UNet(3c)                  | 86.91      | 27.59 | 49.95     | 48.72 | 82.82      | 79.12 | 33.04         | 32.62 | 84.17          | 82.71 | 50.37      | 49.70 | 69.53  | 64.54      | 75.59 | 66.93 | 53.41      | 60.10 |
| UNet(4c)                  | 83.94      | 20.74 | 40.57     | 40.40 | 77.96      | 69.36 | 1.06          | 1.05  | 72.83          | 71.81 | 21.19      | 20.94 | 55.75  | 49.59      | 60.30 | 54.42 | 37.38      | 44.60 |
| UNet(6c)                  | 86.15      | 22.92 | 50.14     | 49.34 | 74.60      | 67.76 | 10.28         | 10.20 | 75.84          | 74.51 | 33.72      | 33.52 | 60.20  | 55.12      | 66.26 | 60.18 | 43.04      | 49.50 |
| RTNet(4c)                 | 77.36      | 28.18 | 62.64     | 60.46 | 92.03      | 81.62 | 40.73         | 36.86 | 75.19          | 74.49 | 52.15      | 51.02 | 71.77  | 66.68      | 76.78 | 71.33 | 55.44      | 61.37 |
| MSNet(4c)                 | 85.29      | 29.45 | 56.02     | 55.28 | 88.47      | 83.17 | 23.44         | 23.27 | 85.84          | 84.35 | 53.77      | 53.14 | 72.50  | 65.47      | 77.03 | 70.78 | 54.78      | 63.09 |
| MSNet(6c)                 | 85.34      | 29.71 | 61.55     | 60.68 | 88.97      | 83.64 | 29.43         | 29.15 | 85.63          | 84.21 | 51.78      | 51.21 | 72.82  | 67.12      | 78.00 | 72.15 | 56.43      | 63.44 |

Table 7. Test results of different model structures in GID dataset.
are used to generate more Multi-spectral index layers will theoretically help to improve the model segmentation effect, and also enable weakly supervised or even unsupervised segmentation of conventional features such as impervious surfaces, vegetation, and water (Waser et al. 2021; Xun et al. 2021), but caution should be exercised to increase the autocorrelation that may result from band data, and to avoid the curse of dimensionality similar to hyperspectral remote sensing. In the subsequent work, we will carry out the research focusing on the semantic segmentation by fusing the spectral information of remote sensing images, designing the end-to-end neural network structure, and integrating the spectral information and geological knowledge into the network framework, which will be conducive to solve the hard nut to crack for feature classification of natural geographical elements.

6. Conclusion

This paper mainly introduces our proposed multi-spectral semantic segmentation network MSNet for remote sensing images. Through experimental validation and test in two public datasets, MSNet has competitive performance in model prediction accuracy, training time-consuming and inference speed compared with the same type of semantic segmentation models; The contribution of the spectral feature information provided by NIR, DSM, NDVI, and NDWI to the semantic segmentation accuracy when fused with visible light is demonstrated. A method is proposed for the dynamic synthesis of multichannel remote sensing image datasets, loading multi-spectral image data with simultaneous band calculation to generate spectral indices NDVI and NDWI, and the experimental results of MSNet show that the overall segmentation accuracy is improved, especially the extraction effect of vegetation and artificial objects is most obvious. It further proves the advantages of spectral features in the semantic segmentation task of remote sensing images.

The original intention of our proposed MSNet is to make full use of all the band data of multispectral remote sensing images, and we try to increase the spectral feature information by obtaining spectral indices through band calculation and dynamically synthesizing multichannel remote sensing image datasets, and we have achieved excellent results, proving that the addition of spectral information helps to improve the semantic segmentation accuracy of multispectral remote sensing images. At this stage, the intelligent interpretation of remote sensing images based on deep learning still widely adopts the mature neural network model oriented to digital pictures, and it is urgent to design deep learning neural networks dedicated to the field of remote sensing. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

Acknowledgements

The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by National Key Research and Development Program of China (No. 2021YFB3900603).

ORCID

Wen Zhang i http://orcid.org/0000-0001-5576-5973

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

2D Semantic Labeling - Potsdam [WWW Document]. n.d. Accessed 10 January 2021. https://www2.isprs.org/commisions/comm2/wg4/benchmark/2d-sem-label-potsdam/

Audebert, N., B. Le Saux, and S. Lefèvre. 2017. “Semantic Segmentation of Earth Observation Data Using Multimodal and Multi-scale Deep Networks.” In Computer Vision – ACCV 2016, Lecture Notes in Computer Science, edited by S.-H. Lai, V. Lepetit, K. Nishino, and Y. Sato, 180–196. Cham: Springer International Publishing. doi:10.1007/978-3-319-54181-5_12.

Badrinarayanan, V., A. Kendall, and R. Cipolla. 2017. “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (12): 2481–2495. doi:10.1109/TPAMI.2016.2644615.
Benjdira, B., Y. Bazi, A. Koubaa, and K. Ouni. 2019. “Unsupervised Domain Adaptation Using Generative Adversarial Networks for Semantic Segmentation of Aerial Images.” Remote Sensing 11 (11): 1369. doi:10.3390/rs11111369.

Carvalho, O. L. F. D., O. A. de Carvalho Júnior, A. O. D. Albuquerque, P. P. D. Bem, C. R. Silva, P. H. G. Ferreira, R. D. S. D. Moura, R. A. T. Gomes, R. F. Guimarães, and D. L. Borges. 2021. “Instance Segmentation for Large, Multi-Channel Remote Sensing Imagery Using Mask-RCNN and a Mosaicking Approach.” Remote Sensing 13 (1): 39. doi:10.3390/rs13010039.

Chen, L.-C., G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. 2018. “DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs.” IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (4): 834–848. doi:10.1109/TPAMI.2017.2699184.

Chollet, F. 2017. “Xception: Deep Learning with Depthwise Separable Convolutions,” in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800–1807. doi:10.1109/CVPR.2017.195.

Davis, J., and M. Goadrich. 2006. “The Relationship between Precision-Recall and ROC Curves,” in: Proceedings of the 23rd International Conference on Machine Learning - ICML ’06. Presented at the 23rd international conference, ACM Press, Pittsburgh, Pennsylvania, pp. 233–240. doi:10.1145/1143844.1143874.

Gao, B. 1996. “NDWI — A Normalized Difference Water Index for Remote Sensing of Vegetation Liquid Water from Space.” Remote Sensing of Environment 58 (3): 257–266. doi:10.1016/S0034-4257(96)00067-3.

GDAL — GDAL Documentation [WWW Document]. n.d. Accessed 13 November 2021. https://gdal.org/.

Guo, Z., Q. Chen, G. Wu, Y. Xu, R. Shibasaki, and X. Shao. 2017. “Village Building Identification Based on Ensemble Convolutional Neural Networks.” Sensors 17 (11): 2487. doi:10.3390/s17112487.

Guo, R., J. Liu, N. Li, S. Liu, F. Chen, B. Cheng, J. Duan, X. Li, and C. Ma. 2018. “Pixel-Wise Classification Method for High Resolution Remote Sensing Imagery Using Deep Neural Networks.” ISPRS Journal of Photogrammetry and Remote Sensing, Deep Learning RS Data 145: 60–77. doi:10.1016/j.isprsjprs.2018.04.014.

Hazirbas, C., L. Ma, C. Domokos, and D. Cremers. 2017. “FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture.” In Computer Vision – ACCV 2016, Lecture Notes in Computer Science, edited by S.-H. Lai, V. Lepetit, K. Nishino, and Y. Sato, 213–228. Cham: Springer International Publishing. doi:10.1007/978-3-319-54181-5_14.

He, K., X. Zhang, S. Ren, and J. Sun. 2016. “Deep Residual Learning for Image Recognition,” in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, pp. 770–778. doi:10.1109/CVPR.2016.90.

He, K., R. Girshick, and P. Dollar. 2019. “Rethinking ImageNet Pre-Training,” in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4971–4976. doi:10.1109/ICCV.2019.00502.

Huang, H., X. Wu, X. Wang, and W. Sillam. 2015. “Maximum F1-Score Discriminative Training Criterion for Automatic Mispronunciation Detection.” IEEE/ACM Transactions on Audio, Speech and Language Processing 23 (4): 787–797. doi:10.1109/TASLP.2015.2409733.

Huang, Z., X. Wang, Y. Wei, L. Huang, H. Shi, W. Liu, and T. S. Huang. 2020. “CCNet: Crisis-Cross Attention for Semantic Segmentation.” arXiv:1811.11721 [cs].

Iwashita, Y., K. Nakashima, A. Stoica, and R. Kurazume. 2019. “TU-Net and TDeepLab: Deep Learning-Based Terrain Classification Robust to Illumination Changes, Combining Visible and Thermal Imagery,” in: 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR). Presented at the 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 280–285. doi:10.1109/MIPR.2019.00057.

Kampffmeyer, M., A.-B. Salberg, and R. Jenssen. 2016. “Semantic Segmentation of Small Objects and Modeling of Uncertainty in Urban Remote Sensing Images Using Deep Convolutional Neural Networks,” in: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 680–688. doi:10.1109/CVPRW.2016.90.

Kemker, R., C. Salvaggio, and C. Kanan. 2018. “Algorithms for Semantic Segmentation of Multispectral Remote Sensing Imagery Using Deep Learning.” ISPRS Journal of Photogrammetry and Remote Sensing, Deep Learning RS Data 145: 60–77. doi:10.1016/j.isprsjprs.2018.04.014.

Khosboreh-Masouleh, M., and R. Shahhosseini. 2021. “A Deep Multi-Modal Learning Method and A New RGB-Depth Data Set for Building Roof Extraction.” Photogrammetric Engineering and Remote Sensing 87 (10): 759–766. doi:10.14358/PERS.21-0007R2.

Konapala, G., S. V. Kumar, and S. Khalique Ahmad. 2021. “Exploring Sentinel-1 and Sentinel-2 Diversity for Flood Inundation Mapping Using Deep Learning.” ISPRS Journal of Photogrammetry and Remote Sensing 180: 163–173. doi:10.1016/j.isprsjprs.2021.08.016.
Li, Q., Q. Wang, and X. Li. 2021. “Exploring the Relationship between 2D/3D Convolution for Hyperspectral Image Super-Resolution.” *IEEE Transactions on Geoscience and Remote Sensing* 59 (10): 8693–8703. doi:10.1109/TGRS.2020.3047363.

Li, R., S. Zheng, C. Zhang, C. Duan, J. Su, L. Wang, and P. M. Atkinson. 2021. “Multattention Network for Semantic Segmentation of Fine-Resolution Semantic Imaging.” *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–13. doi:10.1109/TGRS.2021.3093977.

Lin, D., R. Zhang, Y. Ji, P. Li, and H. Huang. 2020. “SCN: Switchable Context Network for Semantic Segmentation of RGB-D Images.” *IEEE Transactions on Cybernetics* 50 (3): 1120–1131. doi:10.1109/TCYB.2018.2885062.

Liu, Y., B. Fan, L. Wang, J. Bai, S. Xiang, and C. Pan. 2018. “Semantic Labeling in Very High Resolution Images via a self-cascaded Convolutional Neural Network.” *ISPRS Journal of Photogrammetry and Remote Sensing* 145: 78–95. doi:10.1016/j.isprsjprs.2017.12.007.

Liu, X., L. Jiao, L. Li, X. Tang, and Y. Guo. 2021. “Deep multi-level Fusion Network for multi-source Image pixel-wise Classification.” *Knowledge-Based Systems* 221: 106921. doi:10.1016/j.knosys.2021.106921.

Long, J., E. Shelhamer, and T. Darrell. 2015. “Fully Convolutional Networks for Semantic Segmentation, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).” Presented at the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Boston, MA, USA, pp. 3431–3440. doi:10.1109/CVPR.2015.7298965.

Loshchilov, I., and F. Hutter. 2018. *Fixing Weight Decay Regularization in Adam*. Marmanis, D., K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla. 2018. “Classification with an Edge: Improving Semantic Image Segmentation with Boundary Detection.” *ISPRS Journal of Photogrammetry and Remote Sensing* 135: 158–172. doi:10.1016/j.isprsjprs.2017.11.009.

Pan, B., Z. Shi, X. Xu, T. Shi, N. Zhang, and X. Zhu. 2019a. “CoinNet: Copy Initialization Network for Multispectral Imagery Semantic Segmentation.” *IEEE Geoscience and Remote Sensing Letters* 16 (5): 816–820. doi:10.1109/LGRS.2018.2880756.

Pan, X., F. Yang, L. Gao, Z. Chen, B. Zhang, H. Fan, and J. Ren. 2019b. “Building Extraction from High-Resolution Aerial Imagery Using a Generative Adversarial Network with Spatial and Channel Attention Mechanisms.” *Remote Sensing* 11 (8): 917. doi:10.3390/rs11080917.

Panboonyuen, T., K. Jitkajornwanch, S. Lawawirojwong, P. Srestasathier, and P. Vateekul. 2019. “Semantic Segmentation on Remotely Sensed Images Using an Enhanced Global Convolutional Network with Channel Attention and Domain Specific Transfer Learning (Preprint).” *Engineering*. doi:10.20944/preprints201812.0090v3.

Qin, X., Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand. 2020. “U^2-Net: Going Deeper with Nested U-Structure for Salient Object Detection.” *Pattern Recognition* 106: 107404. doi:10.1016/j.patcog.2020.107404.

Romera, E., J. M. Alvarez, L. M. Bergasa, and R. Arroyo. 2018. “ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation.” *IEEE Transactions on Intelligent Transportation Systems* 19 (1): 263–272. doi:10.1109/TITS.2017.2750080.

Ronneberger, O., P. Fischer, and T. Brox. 2015. “U-Net: Convolutional Networks for Biomedical Image Segmentation.” In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, edited by N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, 234–241. Cham: Springer International Publishing. doi:10.1007/978-3-319-24574-4_28.

Rouse, J. W., R. H. Haas, J. A. Schell, and D. W. Deering. 1974. “Monitoring Vegetation Systems in the Great Plains with ERTS.”

Roy, S. K., G. Krishna, S. R. Dubey, and B. B. Chaudhuri. 2020. “HybridSN: Exploring 3-D–2-D CNN Feature Hierarchy for Hyperspectral Image Classification.” *IEEE Geoscience and Remote Sensing Letters* 17 (2): 277–281. doi:10.1109/LGRS.2019.2918719.

Schuegraf, P., and K. Bittner. 2019. “Automatic Building Footprint Extraction from Multi-Resolution Remote Sensing Images Using a Hybrid FCN.” *Isprs International Journal of Geo-Information* 8 (4): 191. doi:10.3390/ijgi8040191.

Shi, W., J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. 2016. “Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).” Presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1874–1883. doi:10.1109/CVPR.2016.207.

Song, S., S. P. Lichtenberg, and J. Xiao. 2015. “SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).” Presented at the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 567–576. doi:10.1109/CVPR.2015.7298655.

Sun, Y., W. ZuO, and M. Liu. 2019. “RTfNet: RGB-Thermal Fusion Network for Semantic Segmentation of Urban Scenes.” *IEEE Robotics and Automation Letters* 4 (3): 2576–2583. doi:10.1109/LRA.2019.2904733.

Tong, X.-Y., G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang. 2020. “Land-cover Classification with high-resolution Remote Sensing Images Using Transferable Deep Models.” *Remote Sensing of Environment* 237: 111322. doi:10.1016/j.rse.2019.111322.

Trias-Sanz, R., G. Stamon, and J. Louchet. 2008. “Using Colour, Texture, and Hierarchical Segmentation for high-resolution Remote Sensing.” *ISPRS Journal of Photogrammetry and Remote Sensing* 63 (2): 156–168. doi:10.1016/j.isprsjprs.2007.08.005.

Waldner, F., and F. I. Diakogiannis. 2020. “Deep Learning on Edge: Extracting Field Boundaries from Satellite Images with a Convolutional Neural Network.” *Remote Sensing of Environment* 245: 111741. doi:10.1016/j.rse.2020.111741.

Wang, J., Z. Wang, D. Tao, S. See, and G. Wang. 2016. “Learning Common and Specific Features for RGB-D Semantic Segmentation with Deconvolutional Networks.” In *Computer Vision – ECCV 2016*, Lecture Notes in Computer
Wang, P., P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell. 2018. "Understanding Convolution for Semantic Segmentation, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)." Presented at the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1451–1460. doi:10.1109/WACV.2018.00163.

Wang, H., R. Fan, Y. Sun, and M. Liu. 2021. "Dynamic Fusion Module Evolves Drivable Area and Road Anomaly Detection: A Benchmark and Algorithms." IEEE Transactions on Cybernetics 1–11. doi:10.1109/TCYB.2021.3064089.

Waser, L. T., M. Rüetschi, A. Psomas, D. Small, and N. Rehush. 2021. "Mapping Dominant Leaf Type Based on Combined Sentinel-1/-2 Data – Challenges for Mountainous Countries." ISPRS Journal of Photogrammetry and Remote Sensing 180: 209–226. doi:10.1016/j.isprsjprs.2021.08.017.

Xiao, T., Y. Liu, B. Zhou, Y. Jiang, and J. Sun. 2018. "Unified Perceptual Parsing for Scene Understanding." arXiv:1807.10221 [cs].

Xun, L., J. Zhang, D. Cao, S. Yang, and F. Yao. 2021. "A Novel Cotton Mapping Index Combining Sentinel-1 SAR and Sentinel-2 Multispectral Imagery." ISPRS Journal of Photogrammetry and Remote Sensing 181: 148–166. doi:10.1016/j.isprsjprs.2021.08.021.

Yuan, X., J. Shi, and L. Gu. 2021. "A Review of Deep Learning Methods for Semantic Segmentation of Remote Sensing Imagery." Expert Systems with Applications 169: 114417. doi:10.1016/j.eswa.2020.114417.

Zhao, H., J. Shi, X. Qi, X. Wang, and J. Jia. 2017. "Pyramid Scene Parsing Network, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)." Presented at the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, pp. 6230–6239. doi:10.1109/CVPR.2017.660.