ISPRS Journal of Photogrammetry and Remote Sensing 168 (2020) 182–196

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

NDVI-Net: A fusion network for generating high-resolution normalized difference vegetation index in remote sensing

Hao Zhang [a], Jiayi Ma [a], Chen Chen [b], Xin Tian [a,*]

[a] Electronic Information School, Wuhan University, Wuhan, 430072, China
[b] Department of Electrical and Computer Engineering, University of North Carolina at Charlotte, NC, 28223, USA

ARTICLE INFO

Keywords:
NDVI
Image fusion
Attention mechanism
Deep learning
Pan-sharpening

ABSTRACT

Normalized difference vegetation index (NDVI), derived from multi-spectral (MS) images, is a metric widely used to evaluate the growth status of vegetation in remote sensing. Existing methods for generating high-resolution (HR) NDVI are typically based on pan-sharpening, which often result in huge errors even in case of tiny spectral distortions. To overcome this challenge, from a novel perspective, this paper introduces an HR vegetation index (HRVI) to realize direct fusion with a low-resolution NDVI rather than pan-sharpening an HRMS image. In particular, we propose a two-branch network based on the multi-scale and attention mechanism, termed as NDVI-Net, to obtain the HRNDVI with small distortion. In our network, the multi-scale channel enhancement blocks are used in both NDVI and HRVI branches, in which multi-scale convolution is used to capture structural information with different reception fields and channel attention mechanism is adopted to perform feature selection. Meanwhile, the spatial features are injected unidirectionally from the HRVI into NDVI branches, so as to further improve the quality of features in the NDVI branch. Subsequently, the spatial intensify block is adopted only in the NDVI branch to implement selective enhancement for the previously obtained features along the spatial position, strengthening the retention of local detail features. Finally, HRNDVI is reconstructed based on the high-representation NDVI features, which contains clear texture details and precise intensity. Experimental results demonstrate the significant advantage of our method over the current state-of-the-art in terms of both subjective visual effect and quantitative metrics. Moreover, we apply the HRNDVI generated by our method to vegetation detection and enhancement, and land cover mapping in remote sensing, which can achieve the best performance.

1. Introduction

Normalized difference vegetation index (NDVI) is proposed in Tucker (1979) to assess the level of green vegetation, which is calculated from the near-infrared (NIR) band and the red (R) band in the multi-spectral (MS) image according to the following equation:

$$NDVI = \frac{NIR - R}{NIR + R}. \quad (1)$$

Because of its excellent vegetation characterization performance, NDVI has become one of the most important indicators in the field of remote sensing (Carlson and Ripley, 1997; Zhu and Liu, 2015; Yang et al., 2012). However, it is difficult for remote sensing satellites to obtain high-resolution (HR) MS images, which is caused by the characteristics of MS sensors. More specifically, the spectroscopic/filter mechanism in MS sensors requires a large instantaneous field of view (IFOV) to meet the requirement of signal-to-noise ratio, which means it reduces the spatial resolution while ensuring spectral richness of the resulting image. This restriction also indirectly leads to the low-resolution (LR) of NDVI, which largely limits the accuracy of subsequent applications, such as vegetation detection (Zhang et al., 2020b). Therefore, it is desirable to develop a technique to generate HRNDVI.

The existing methods for obtaining HRNDVI are based on pan-sharpening (Wang et al., 2016; Rahmani et al., 2010; Aiazzi et al., 2013). In particular, these methods first fuse the HR panchromatic (PAN) image and the LRMS image to generate the HRMS image, and then calculate HRNDVI from the R and NIR bands of the HRMS. However, pan-sharpening is very difficult to generate precise HRMS. Most pan-sharpening methods generally follow the assumption: the intensity/gradient of the PAN image is a linear combination of the intensity/gradient of multiple channels in the MS image. Unfortunately, the problem of accurately solving linear combination coefficients has

* Corresponding author.
E-mail addresses: zhpersonalbox@gmail.com (H. Zhang), jyma2010@gmail.com (J. Ma), chenchen870713@gmail.com (C. Chen), xin.tian@whu.edu.cn (X. Tian).

https://doi.org/10.1016/j.isprsjprs.2020.08.010
Received 26 April 2020; Received in revised form 19 July 2020; Accepted 11 August 2020
Available online 23 August 2020

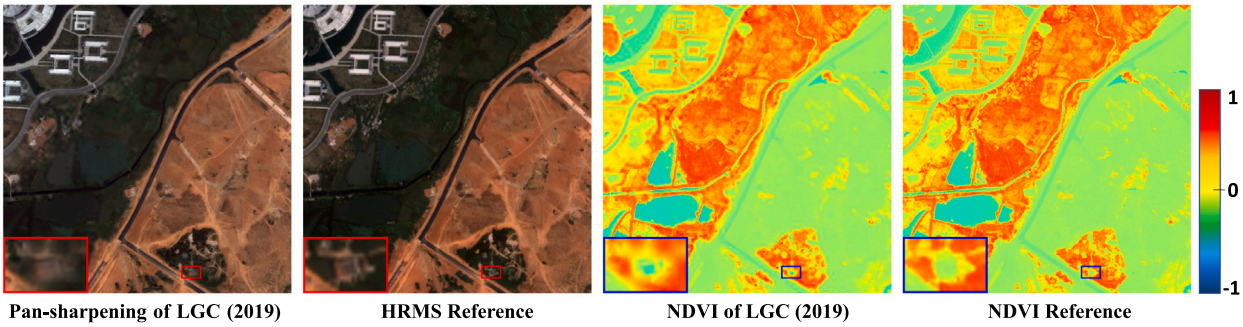

Fig. 1. Small spectral distortion causes huge NDVI errors. From left to right: HRMS generated by LGC (Fu et al., 2019), HRMS reference, NDVI of LGC and NDVI reference.

not yet been solved, which leads to more or less distortion existing in the obtained HRMS.

Further, the values of NIR and R are very small (*i.e.*, close to 0) in some regions. In this case, the distortion in HRMS generated by pan-sharpening will be further amplified by calculating NDVI through Eq. (1). In other words, a small spectral distortion will cause a huge NDVI error, which is harmful to subsequent applications. Similarly, Johnson (2014) also pointed out that the NDVI would suffer from some spatial information loss due to pan-sharpening. To illustrate this problem more intuitively, a latest pan-sharpening method, local gradient constraints (LGC) (Fu et al., 2019), is used as an example in Fig. 1. It can be clearly seen that, compared with the ground truth, there is slight spectral distortion in the pan-sharpening result generated by LGC (Fu et al., 2019), while large errors occur in the calculated NDVI. This error is not only reflected in the intensity (vegetation coverage), but also in the details of the texture (the boundary between vegetation), which is very detrimental to the application of agricultural remote sensing.

To address the above mentioned limitation, in this paper we propose to implement the fusion directly on the NDVI image for generating HRNDVI, rather than pan-sharpening an HRMS image, so as to avoid the error amplification effect of Eq. (1). Nevertheless, there are also great challenges in direct fusion. First of all, it is difficult to define the source data fused with LRNDVI. Specifically, the source data should contain rich spatial texture information to compensate for LRNDVI. Meanwhile, the spatial texture information should be as similar as possible to NDVI, that is, the new source data should have a similar physical meaning to NDVI. Second, there are a lot of complex texture details in NDVI. Some of them are artificial boundaries between vegetation, such as roads and buildings, and others are transition boundaries between areas with deep vegetation coverage and areas with shallow coverage. Preserving these complex and tiny texture details is very challenging.

To overcome the above challenges, we design a novel two-branch network based on the multi-scale and attention mechanism for NDVI fusion, which can generate the HRNDVI with small distortion, termed as *NDVI-Net*. In our method, the above considerations are solved from two aspects.

On the one hand, we introduce the HR vegetation index (HRVI) (Tu et al., 2009) and modify it to be the source data fused with LRNDVI. Concretely, the HRVI is defined as:

$$HRVI = \frac{PAN - R\uparrow}{PAN + R\uparrow}, \tag{2}$$

where the PAN is the HRPAN, R refers to the red band of the LRMS and $\uparrow$ indicates the upsampling function of bicubic. Obviously, the definition of HRVI is similar to that of NDVI. The difference is that the HRPAN image is introduced into the definition of HRVI, so as to contains rich spatial texture information. A visual example isprovided in Fig. 2. It can be observed that the HRVI has similar but clearer textures compared with the NDVI, which can therefore provide spatial information for the reconstruction of HRNDVI. Note that in the original definition of HRVI in Tu et al. (2009), R is replaced with the average of the red, green and blue bands of the LRMS image. The newly defined HRVI in Eq. (2) can introduce spatial information of the PAN image while ensuring its texture structure to be as close as possible to that in the NDVI, thus reducing the difficulty of accurate texture reconstruction. We will show the advantages of our new definition in the experimental section.

On the other hand, we design a specific network to preserve the tiny and complex textures. It is a two-branch network, namely the NDVI branch and HRVI branch, to extract features from the LRNDVI and HRVI and reconstruct HRNDVI. The multi-scale channel enhancement block is used in these two branches at the first stage. In this block, we use convolutions of different scales for feature extraction, because multiple receptive fields can allow more structural information to be contained in the extracted features. Then, the channel attention mechanism selectively enhances more important features after each multi-scale convolution according to the fusion objective. In this process, the spatial features filtered at each layer in the HRVI branch are unidirectionally injected into the NDVI branch to improve the spatial structure quality of features. At the second stage, we use the spatial intensify block to selectively weight the features generated by the previous NDVI-branch network along the pixel position, which can further enhance the retention of feature information with small details. Finally, the high-quality HRNDVI can be reconstructed from the features with high expression ability, which contains clear and accurate texture details.

To intuitively demonstrate the advantages of our method over pan-sharpening-based methods, we provide a typical example of our fused result with comparison to two state-of-the-art methods, *i.e.*, the generalized Laplacian pyramid-based method MTF-GLP (Aiazzi et al., 2006) and deep learning-based method PNN (Masi et al., 2016). MTF-GLP adopts the modulation transfer functions of the multi-spectral scanner to design the generalized Laplacian pyramid reduction filter, so as to realize the spatial injection of pan-sharpening. While PNN trains the neural network under the supervision of reference images to realize pan-sharpening. The results are shown in the top row of Fig. 3. Clearly, our NDVI result is more similar to the reference image. In terms of intensity distribution, our NDVI-Net can provide a relatively more accurate result, which maintains vegetation growth status better. In addition, our method maintains the texture details of NDVI more clearly. It can be seen from the highlighted regions that the dividing line between vegetation is almost the same in our result and the reference, while both MTF-GLP and PNN obscure them.

In addition, the HRNDVI generated by our proposed method can be further applied to the vegetation detection and enhancement of the HRMS image. The results of different methods are provided in the bottom row of Fig. 3. Because the accuracy of HRNDVI obtained by our method is higher than other methods, the effect of enhancing vegetation is the best, especially the preservation of texture between vegetation. As a result, when detecting and enhancing vegetation in the HRMS image, the NDVI generated by our NDVI-Net can be used instead of that generated by pan-sharpening, which can greatly improve the accuracy of vegetation detection and enhancement.

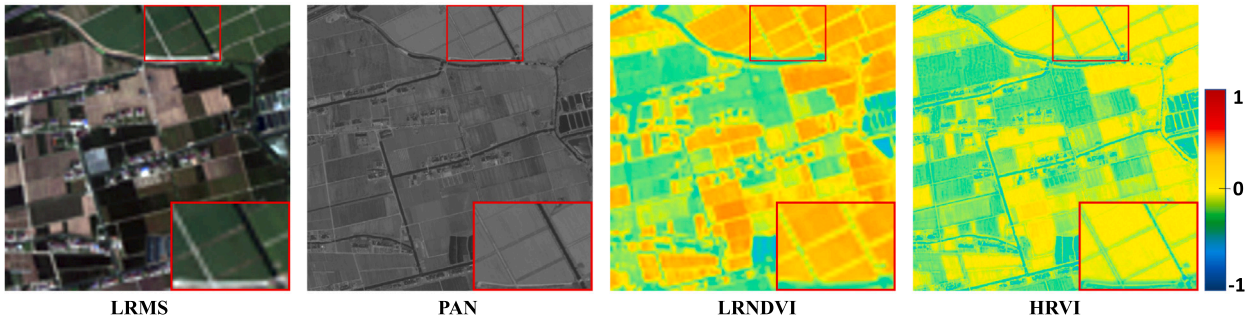

Fig. 2. Visualization of NDVI and HRVI. From left to right: LRMS image (upsampled), PAN image, LRNDVI (upsampled) and HRVI. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

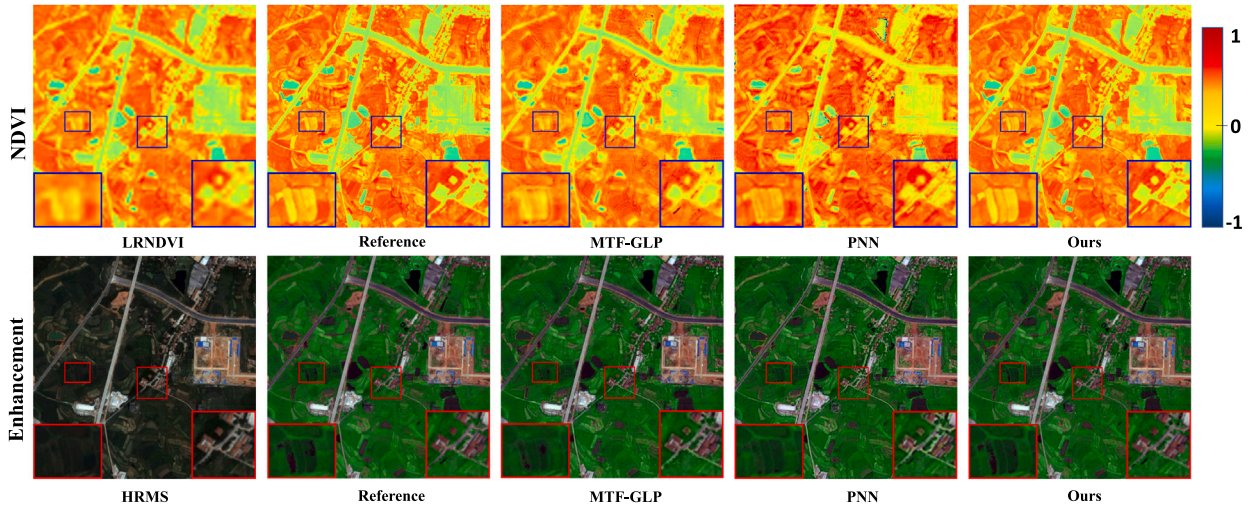

Fig. 3. Illustration of the characteristic of our method. From left to right in the top row: the LRNDVI, reference NDVI, results of MTF-GLP (Aiazzi et al., 2006), PNN (Masi et al., 2016) and our proposed NDVI-Net. From left to right in the bottom row: the HRMS and the vegetation enhancement results based on the corresponding NDVI results.

The major contributions of this paper are summarized as follows. First, we define a new HRVI to achieve direct fusion with NDVI, which can provide spatial texture information for HRNDVI reconstruction, thereby making it possible to generate HRNDVI with clear texture details. Second, we propose a novel two-branch network based on the multi-scale and attention mechanism to realize NDVI fusion, which can generate HRNDVI with clearer texture details and more precise intensity than the traditional pan-sharpening-based methods. Third, the multi-scale channel enhancement block and spatial intensify block are designed, which strengthen the preservation of tiny texture details. Moreover, the unidirectional injection of spatial information between the two branches is also a valuable means of feature quality improvement. Fourth, we also apply our method to NDVI-based vegetation detection and enhancement, and land cover mapping, in which our method can achieve the most consistent results with the reference.

The remainder of this paper is organized as follows. Section 2 describes some related work, including an overview of existing pan-sharpening methods and attention mechanisms. In Section 3, we describe our method in detail, including the overview of the framework, loss functions, and network architecture design. In Section 4, we give the detailed experimental settings and compare our method with several state-of-the-art methods qualitatively and quantitatively on publicly available datasets. In addition, we also carry out the validation of HRVI definition, ablation experiments, visualization of space injection, generalization experiment, experiment of vegetation detection and enhancement, and experiment of land cover mapping in this section. Conclusions are given in Section 5.

2. Related work

This section describes the background and existing works that are most related to our research, including the development of pan-sharpening methods and attention mechanisms.

2.1. Pan-sharpening methods

Pan-sharpening is the most commonly used strategy for obtaining the HRMS image and then calculating the HRNDVI (Duran et al., 2017; Tian et al., 2020). It aims to preserve geometric texture details of the HRPAN image and spectral information of the LRMS image. However, PAN is a single-channel image and MS is a multi-channel image, which makes it difficult to define the correspondence between texture or intensity of them. Most pan-sharpening methods follow the assumption: the PAN image (or its gradient) can be modeled as a linear combination among all bands (or their gradients) of the HRMS image, which can be formalized as:

$$PAN = \sum_{b=1}^{n} \omega_b \cdot MS_b + \varepsilon_1, \tag{3}$$

$$\nabla PAN = \sum_{b=1}^{n} \alpha_b \cdot \nabla MS_b + \varepsilon_2, \tag{4}$$

where PAN and MS represent the HRPAN image and the HRMS image, b is the index of the spectral band, n is the total number of spectral bands in HRMS, and ∇ is the gradient operator. In addition, $\omega_{(\cdot)}$ and $\alpha_{(\cdot)}$ indicate coefficients of linear combination, and $\varepsilon_{(\cdot)}$ is the deviation terms. For Eq. (3), many tentative solutions have been

given. In particular, a simple solution is used in the generalized HIS method (Carper et al., 1990), where the same weights of different bands are adopted. Subsequently, Aiazzi et al. (2007) adopted the optimized way to determine these linear combination coefficients. However, these methods tend to cause severe spectral distortion. Because the response characteristics of different sensors mounted on satellites to objects are very different. Specifically, the union of imaging bands of some MS images is not as extensive as the PAN image. This linear combination in the intensity of multi-spectral image often fails to synthesize a good pseudo-panchromatic image, which reduces the intensity fidelity of HRMS.

After realizing this problem, most methods have recently favored following Eq. (4), which ensures the consistency of the high-pass filtered components of PAN image and HRMS image, instead of intensity. Chen et al. (2015) proposed the SIRF to introduce the dynamic gradient sparsity, which copies the PAN image to the same channel number as the MS image and requires them to have the gradient consistency. Similarly, PMGI (Zhang et al., 2020a) requires that each channel of the MS image has the gradient consistency with the PAN image. It is worth noting that this definition is still problematic. Because the PAN image is a wider band imaging, its texture structure is richer than any channel of the MS image. Therefore, it is unreasonable to copy PAN images to multiple channels and then constrain the gradient consistency. A recent work LGC (Fu et al., 2019) innovatively pointed out that the linear weighting in all the above methods is based on the global perspective, which cannot well model the local relationship between MS and PAN. Based on this observation, a variational pan-sharpening with local gradient constraints is proposed, which can provide a relatively accurate spatial preservation.

In recent years, neural networks have promoted the great development of pan-sharpening. Masi et al. (2016) introduced the supervised convolutional neural network to realize pan-sharpening, which can generate promising HRMS. However, the output of the neural network often has the phenomenon that local details are smoothed under the constraint of ℓ_2 loss. The generative adversarial network (GAN) is also a popular technique to solve the image fusion problem (Xu et al., 2020; Ma et al., 2020a, 2019). In particular, PSGAN (Liu et al., 2018) introduces the GAN to the pan-sharpening problem for the first time, and after continuous adversarial games, the generator can produce results with richer textures, but these textures are often fake due to adversarial learning. Pan-GAN (Ma et al., 2020b) further provides an unsupervised framework for pan-sharpening based on GAN, which does not require the ground-truth during network training.

2.2. Attention mechanism

Conceptually, the attention mechanism is a bionic technical, which is inspired by the observation characteristics of animals. Concretely, when the neural network extracts the features of the viewed target, it should assign different weights to the feature map according to the degree of contribution to the current task. Classically, Itti et al. (1998) proposed a visual attention system, which can quickly realize the scene understanding. Subsequently, attention mechanisms were introduced into various visual tasks. The implementation of attention mechanism has various dimensions, which can be at channel, spatial and temporal.

The attention mechanism along the channel dimension is also called channel attention, which is to selectively weight different types of features extracted by different convolution kernels, so as to achieve feature enhancement or suppression. These features can be high-frequency and low-frequency features. The high-frequency feature often reflects the texture structure of an image, which is more important in some fields. Conversely, the low-frequency feature indicates more holistic information such as the intensity distribution characteristics of image, which is needed in certain fields. Differently, the spatial attention focuses on the characteristics of spatial position. The so-called spatial position is from the length and width of an image, which reflects the spatial relationship of the 3D object projected onto the 2D imaging plane. In some cases, certain regions of an image are more important for specific tasks. For example, the target is more important than the background in the detection task. Another example is that in the infrared image, regions with strong thermal radiation are more important than regions with weak thermal radiation. Temporal attention is different from the above two attention mechanisms, which is proposed for time series data. In some cases, the data features of one moment are more important than those of another. For example, the importance of words in natural language processing is different. Also, the contribution of video frames to action understanding is different, in which frames with changes in action status are more important than those still ones.

The development and application of some typical attention mechanisms are given below. Bahdanau et al. (2014) applied attention mechanism in the field of natural language processing, which makes full use of the language context information and improves the performance of machine translation. Luong et al. (2015) further considered the scale of attention and proposed two attention models, namely, global attention and local attention, which are used in machine translation. In recent years, some plug-and-play attention blocks have been proposed one after another, which can be applied to various tasks. Hu et al. (2018) designed a channel attention module called SI block, which performs feature selection along the channel dimension. Woo et al. (2018) proposed the CBAM block, which extends the SI block from just channel dimension to channel plus spatial dimensions. The effectiveness of SE and CBAM blocks has been widely proven. In SCSCN (Ma et al., 2020c), the separated channel-spatial attention is adopted to focus on the edges and high-frequency features of the target to obtain high-quality 3D reconstruction results. Hua et al. (2019) designed a class attention learning layer, which aims at capturing discriminative class-specific features, so as to improve the accuracy of multi-label aerial image classification. In our NDVI-Net, attention mechanisms are used to screen important features and enhance tiny texture details for NDVI.

3. Method

In this section, we give a detailed introduction to our method. We first introduce the overview of the framework, and then give the definition of loss functions. Finally, the detailed structure of the proposed NDVI-Net is provided.

3.1. Overview of the framework

The purpose of our work is to generate HRNDVI, which contains clear spatial texture and has an accurate intensity distribution. The intensity distribution characteristics are not difficult to infer from LRNDVI under the constraint of the reference image, while it is not easy to change the spatial texture from weak to strong. For this observation, we introduce the HRVI defined by Eq. (2) to provide adequate spatial texture information for HRNDVI reconstruction. On this basis, a new two-branch fusion network is designed to reconstruct clear spatial texture and accurate intensity distribution. As shown in Fig. 4, the two branches of NDVI-Net are the NDVI and HRVI branches. The NDVI branch is the main branch used to recover HRNDVI from LRNDVI, in which the intensity distribution features can be obtained from LRNDVI. The role of the HRVI branch is to extract and select spatial texture features from HRVI and then inject them into the NDVI branch, so that the HRNDVI with a reasonable intensity distribution and clear spatial texture can be generated by the NDVI branch.

Concretely, we firstly up-sample the LRNDVI to the same size as the HRVI, which is implemented by transposed convolution. Secondly, the multi-scale channel enhancement block is separately used in both the NDVI branch and the HRVI branch to extract and select important features. For example, the intensity distribution characteristics of LRNDVI and the spatial texture characteristics of HRVI should be extracted and

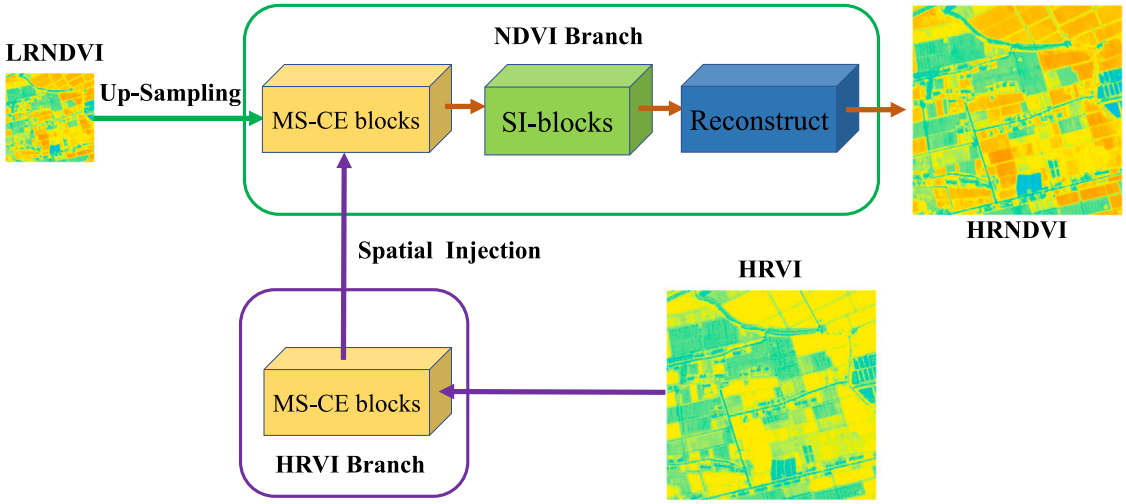

Fig. 4. Overall fusion framework of our NDVI-Net. We adopt NDVI branch and HRVI branch to obtain intensity distribution and spatial texture information, respectively. MS-CE: multi-scale channel enhancement; SI: spatial intensify.

selected, because they are important for reconstructing the HRNDVI. Each multi-scale channel enhancement block consists of a multi-scale convolution and a channel attention. The multi-scale convolution is to extract features from different scale receptive fields, which is conducive to the maintenance of a local structure. The channel attention is to filter the features extracted by different convolution kernels. After each multi-scale channel enhancement block, we inject the features from the HRVI branch into the NDVI branch in a one-way manner, which can improve the spatial information quality of features in the NDVI branch. Thirdly, we adopt the spatial intensify block to selectively weight the features in the NDVI branch along the spatial position, further strengthening the attention for preservation of tiny textures. It is worth noting that we perform feature reuse in each stage in our network to reduce information loss due to convolution. Guided by the specifically designed loss function, our NDVI-Net can reconstruct the preliminary HRNDVI. Finally, we perform the post-processing to reduce the drift of the neural network output. To be specific, we use the traditional method to decompose the preliminary HRNDVI into base layer and detail layer, and then perform histogram specification for the base layer referencing the up-sampled LRNDVI. The final high-quality HRNDVI is obtained by adding the detail layer and the processed base layer, which not only has the intensity distribution similar to ground truth, but also contains fine texture details.

3.2. Loss functions

The loss function is designed based on the intensity distribution and texture details, which consists of an intensity loss term $\mathcal{L}_{\text{int}}$ and a gradient loss term $\mathcal{L}_{\text{grad}}$:

$$\mathcal{L} = \mathcal{L}_{\text{int}} + \gamma \mathcal{L}_{\text{grad}}, \tag{5}$$

where γ is used to balance the two loss terms.

The intensity loss term is used to constrain the intensity distribution of the reconstructed HRNDVI to approximate that of the reference image. In order to reduce the detail smoothing effect caused by regression (Zhao et al., 2016), we use ℓ_1 loss instead of ℓ_2. The intensity loss term $\mathcal{L}_{\text{int}}$ is formalized as:

$$\mathcal{L}_{\text{int}} = \frac{1}{HW} |I_{\text{fused}} - I_{\text{refer}}|, \tag{6}$$

where I_{fused} is the HRNDVI generated by the network, I_{refer} is the ground truth, H and W are the height and width of the image, respectively.

Only using the intensity loss term still inevitably causes some local details to be blurred. In order to preserve the tiny details, we introduce the gradient loss term, in which we use the Sobel operator to find the gradient. It is worth noting that we constrain the gradient consistency in the X and Y directions between the fused NDVI and the reference NDVI, instead of merging the gradient of these two dimensions. In other words, we require their gradients to be numerically equal, and also want them to be in the same direction. We again choose the ℓ_1 loss in the gradient loss term $\mathcal{L}_{\text{grad}}$, which is defined as:

$$\mathcal{L}_{\text{grad}} = \frac{1}{HW} |\nabla_x I_{\text{fused}} - \nabla_x I_{\text{refer}}| + \left|\nabla_y I_{\text{fused}} - \nabla_y I_{\text{refer}}\right|, \tag{7}$$

where ∇_x and ∇_y are Sobel gradient operators in X and Y directions.

3.3. Network architectures

The NDVI-Net we proposed is a two-branch convolutional neural network, which is shown in Fig. 5. First, two transposed convolution layers with the 5×5 convolution kernel are used in the NDVI branch to up-sample the LRNDVI to the same size as HRVI. Second, both NDVI and HRVI branches use four multi-scale channel enhancement blocks to extract and select the required features. The detailed structure of the multi-scale channel enhancement block is shown in the lower right corner of Fig. 5. In particular, three convolution layers with 3×3, 5×5, and 7×7 convolution kernels are first used, and then the output of them is concatenated. Based on the concatenated result, a channel attention map can be generated, then we multiply the concatenated features and the attention map to obtain enhanced features. After each multi-scale channel enhancement block, we unidirectionally inject the features in the HRVI branch into the NDVI branch to improve the spatial quality of the features in the NDVI branch. Third, we adopt four spatial intensify blocks in the NDVI branch to strengthen the preservation of tiny details. The spatial intensify block consists of a convolution layer with 5×5 convolution kernel and spatial attention, which is also shown in the lower right corner of Fig. 5. In detail, based on the features extracted by this convolutional layer, a spatial attention map is generated, and then the spatially enhanced features are obtained by multiplying the feature and the spatial attention map. Finally, two convolutional layers with 5×5 convolution kernel are used to reconstruct the HRNDVI. Except for the last layer, all convolutional layers use the Leaky ReLU as the activation function, while the last convolutional layer adopts the Tanh as the activation function.

4. Experiments

In this section, we verify the performance of our NDVI-Net. Firstly, we introduce the experimental settings including datasets, training

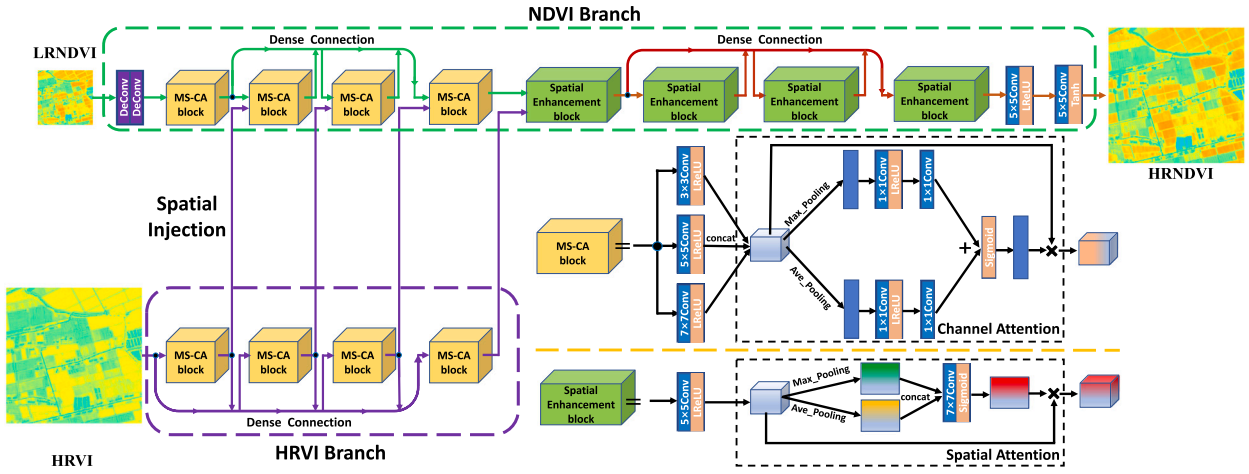

Fig. 5. Network architecture of the proposed NVDI-Net. The structural diagrams of MS-CA block and spatial enhancement block are located in the lower right corner.

details and evaluation metrics. Then, we provide qualitative and quantitative results on two datasets. In addition, we compare different definitions of HRVI to prove the rationality of our modified HRVI in reconstructing HRNDVI. We also perform some ablation experiments to demonstrate the role of the HRVI branch and the effectiveness of the spatial intensify block and multi-scale channel enhancement block. Subsequently, the generalization experiment is provided. Finally, we conduct additional applied experiments, namely vegetation detection and enhancement, and land cover mapping.

4.1. Experimental settings

4.1.1. Datasets

To fully validate the performance of our method, we use the QuickBird and GF-2 datasets for evaluation. The spatial resolutions of high-resolution NDVI in these two satellites are 0.61 m and 1 m. We follow Wald's protocol (Wald et al., 1997) to down-sample the original high-resolution NDVI to low-resolution NDVI and then the high-resolution NDVI is used as the reference. In other words, the spatial resolution of low-resolution NDVI in these two satellites are 2.44 m and 4 m, respectively. On these two datasets, the number of image pairs used for testing is both 25. For training, in order to obtain more training data, we adopt the expansion strategy of tailoring and decomposition. Specifically, for the QuickBird, we crop the rest of data to 43,940 image patch pairs for training; for the GF-2, we crop the rest of images to 60,840 image patch pairs for training. It should be noted that in these training data, the size of LRNDVI is 25×25, and the size of HRVI is 100×100. In addition, the data input into the network conform to the definitions of NDVI and HRVI, which are in the range of $[-1, 1]$.

4.1.2. Training details

The batch size is set to b, and it takes m steps to train one epoch. The total number of training epochs is M. In our experiment, we set $b = 32$, $M = 65$, and m is set as the ratio between the whole number of patches and b. In addition, the ratio of intensity loss term and gradient loss term in the loss function is $1 : 5$, that is to say, γ is set to 5. The parameters in our NDVI-Net are updated by AdamOptimizer. All deep learning-based methods run on the same GPU RTX 2080Ti, while other methods run on the same CPU Intel i7-8750H.

4.1.3. Evaluation metrics

We evaluate the fused results from two aspects, *i.e.*, qualitatively and quantitatively. Qualitative evaluation relies on human subjective visual perception. A good fused result should have similar intensity distribution and fine texture details as the reference HRNDVI. Quantitative evaluation refers to the statistical calculation of fused images, using some statistic values to represent the fusion quality. We select six statistics as objective metrics to measure the fused results, such as the root mean square error (RMSE), gradient magnitude similarity deviations (GMSD) (Xue et al., 2013), structural similarity index measure (SSIM) (Wang and Bovik, 2002), correlation coefficient (CC) (Deshmukh and Bhosale, 2010), visual information fidelity (VIF) (Sheikh and Bovik, 2006) and information fidelity criterion (IFC) (Sheikh et al., 2005). All the six metrics are calculated based on the reference HRNDVI. Concretely, RMSE and GMSD measure the pixel difference and gradient difference between the fused result and the reference HRNDVI, respectively. The smaller the RMSE and GMSD, the better the quality of the fused result. On the contrary, SSIM, CC, VIF and IFC respectively evaluate the structural similarity, correlation, visual fidelity and information fidelity between the generated HRNDVI and reference HRNDVI. The larger these four metrics, the better the image quality.

4.2. Comparative experiments

We evaluate our NDVI-Net with comparison to seven state-of-the-art methods including BDSD (Garzelli et al., 2007), PRACS (Choi et al., 2010), PMGI (Zhang et al., 2020a), MTF-GLP (Aiazzi et al., 2006), PNN (Masi et al., 2016), LGC (Fu et al., 2019) and NTV (Zhang et al., 2020b).

4.2.1. Qualitative comparison

Two typical image pairs from QuickBird and GF-2 are selected to qualitatively demonstrate the characteristics of our proposed method, as shown in Figs. 6 and 8. From these results, we see that our NDVI-Net has clear advantages over other methods. First of all, compared with other methods, our method can maintain a more accurate intensity distribution. The intensity of NDVI represents the growth status of vegetation, so the intensity accuracy of fusion results is very important to vegetation detection. The overall intensity distribution of the NDVI images obtained by our method is closer to that of the reference HRNDVI, while other comparative methods have more or less intensity distortion. Second, our method has high fidelity to local texture details. The local texture details of NDVI are often the basis for the division of vegetation areas, and its fidelity directly determines the accuracy of the division of each area. As highlighted in Fig. 6, our method is effective in maintaining the boundaries between houses, roads and vegetation. In Fig. 8, our method can effectively maintain the texture details of the pond vegetation connection, which cannot be achieved by all other comparative methods.

In addition, we provide residual images between the results of each method and the reference HRNDVI to demonstrate the degree of

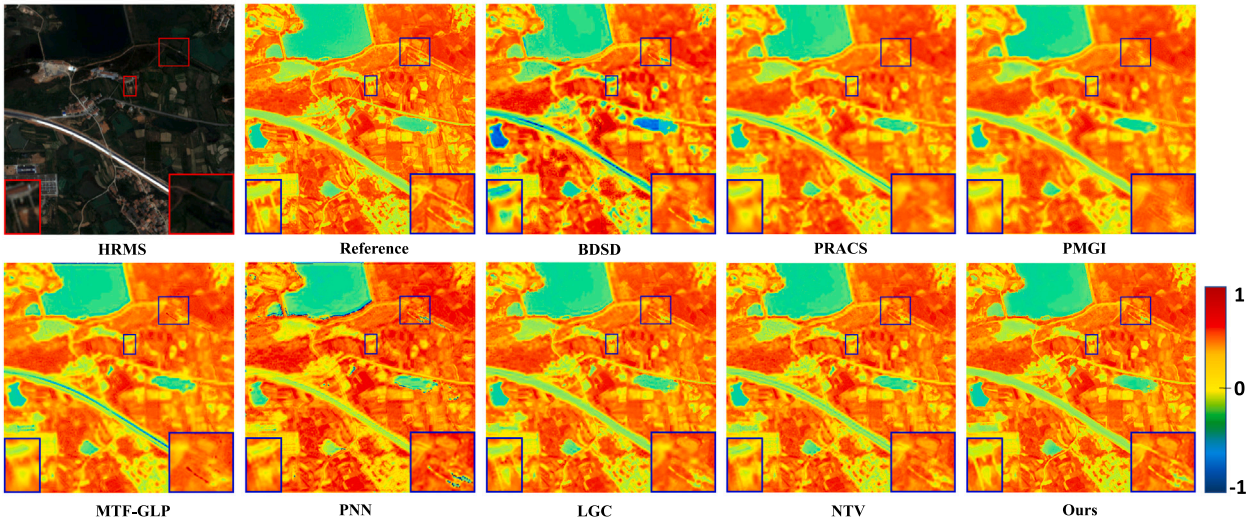

Fig. 6. Qualitative comparison of different methods for NDVI fusion on the data from QuickBird. The images are HRMS image, reference NDVI, fused results of BDSD (Garzelli et al., 2007), PRACS (Choi et al., 2010), PMGI (Zhang et al., 2020a), MTF-GLP (Aiazzi et al., 2006), PNN (Masi et al., 2016), LGC (Fu et al., 2019), NTV (Zhang et al., 2020b) and our NDVI-Net.

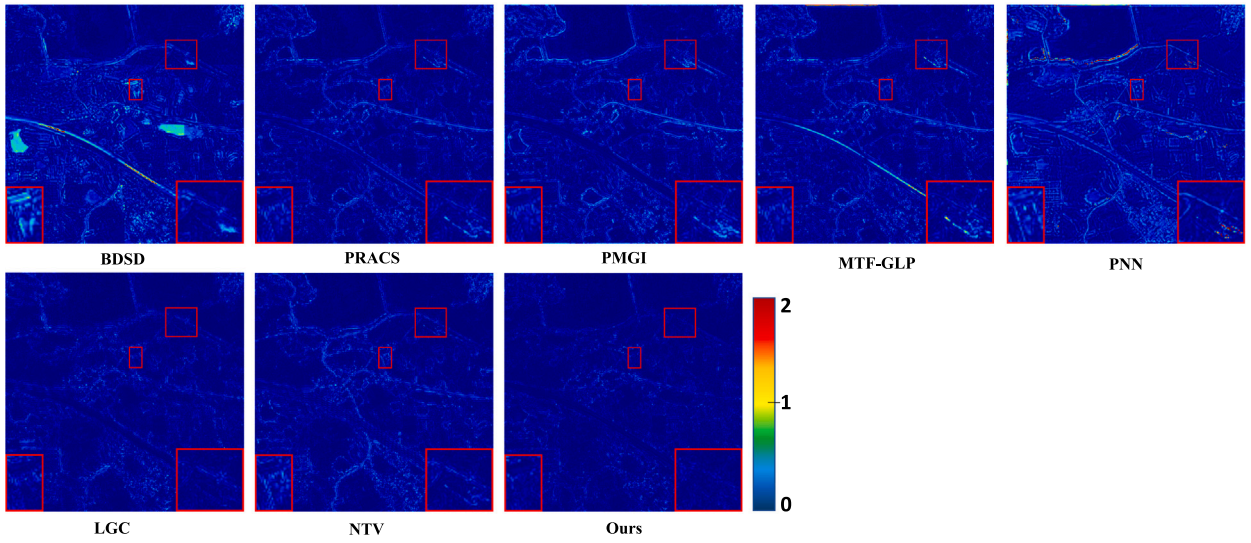

Fig. 7. The residual NDVI by the absolute error between the fused result and reference NDVI in Fig. 6.

distortion, as shown in Figs. 7 and 9. We also highlight the same regions as in Figs. 6 and 8 for comparison. Obviously, the residual image of our method is the darkest compared with other methods, which shows the high fidelity characteristics of our NDVI-Net. In general, our NDVI-Net performs better than other methods in terms of subjective perception, which not only has a more accurate intensity distribution, but also contains richer local texture details.

4.2.2. Quantitative comparison

In order to assess our method more comprehensively, we further provide quantitative comparisons of the seven comparative methods on the two datasets from QuickBird and GF-2. The six metrics RMSE, GMSD, SSIM, CC, VIF and IFC all need the reference HRNDVI to calculate. The statistical results are shown in Tables 1 and 2.

From the results, we see that our NDVI-Net is able to achieve much better average values than all the other competitors on all the six metrics on both datasets. The RMSE and GMSD metrics show that the results of our method have minimal intensity and gradient differences from reference HRNDVI. In addition, SSIM, CC, VIF and IFC show that our NDVI-Net can maintain the structural similarity and correlation degree with reference HRNDVI to the greatest extent, and has the best visual information fidelity. All of them are consistent with the visual perception of our results. Overall, our NDVI-Net performs significantly better than all other methods in objective evaluation.

4.3. Validation of HRVI definition

In this work, we introduce the HRVI (Tu et al., 2009) and modify it to be the source data fused with LRNDVI. Compared with the original HRVI, the modified HRVI has the texture more similar to NDVI, which facilitates the reconstruction of fine textures. To verify this point, a comparative experiment is provided in Fig. 10. It can be seen that our modified HRVI has a distribution similar to the reference NDVI, especially the texture details. On the contrary, the HRVI defined in Tu et al. (2009) is visually different from the reference NDVI, and some tiny textures are quite weak. The corresponding fused results further demonstrate this difference. In the HRNDVI generated by fusing LRNDVI and our modified HRVI, the tiny intervals between vegetation can be well reconstructed. But these intervals are weak in the HRNDVI generated by fusing LRNDVI and the origin HRVI. As a result, our modified HRVI is more reasonable in generating HRNDVI.

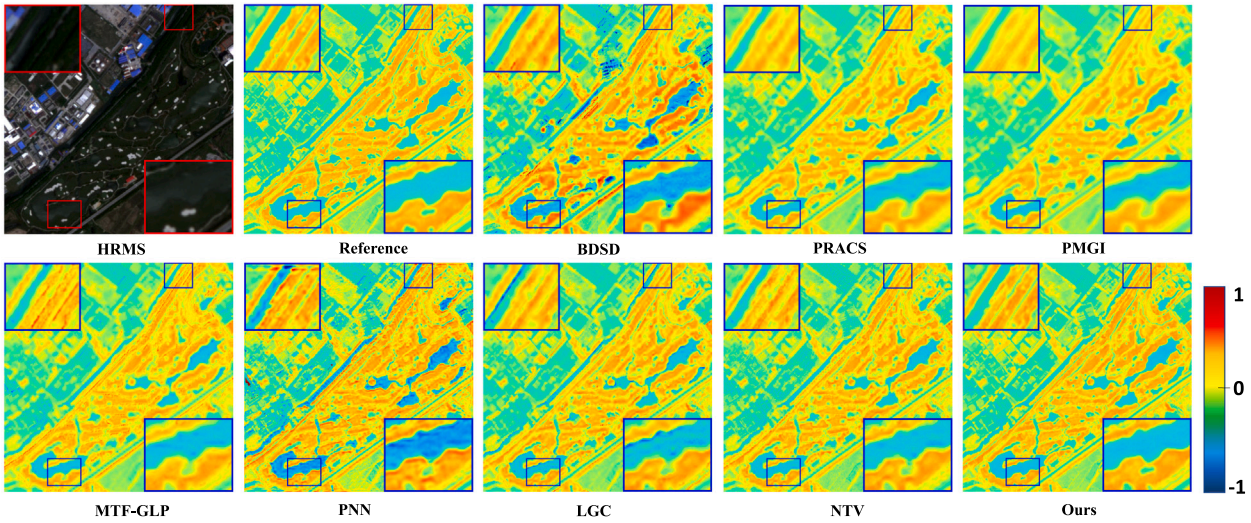

Fig. 8. Qualitative comparison of different methods for NDVI fusion on the data from GF-2. The images are HRMS image, reference NDVI, fused results of BDSD (Garzelli et al., 2007), PRACS (Choi et al., 2010), PMGI (Zhang et al., 2020a), MTF-GLP (Aiazzi et al., 2006), PNN (Masi et al., 2016), LGC (Fu et al., 2019), NTV (Zhang et al., 2020b) and our NDVI-Net.

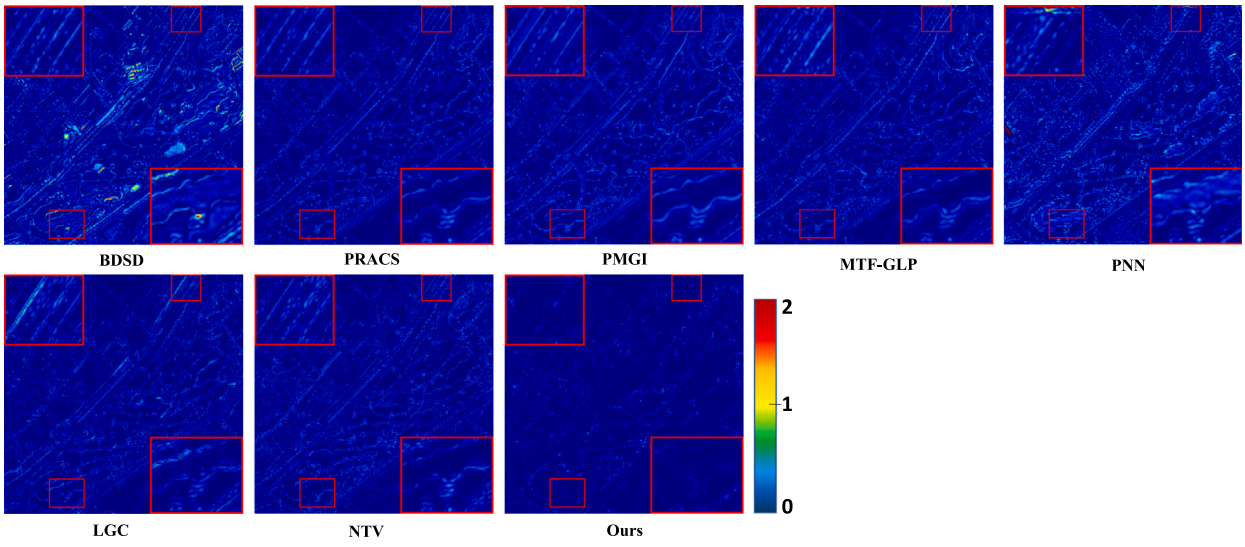

Fig. 9. The residual NDVI by the absolute error between the fused result and reference NDVI in Fig. 8.

4.4. Ablation experiments

In order to verify the effectiveness of the specific designs in this paper, we perform relevant ablation experiments, including the HRVI branch, multi-scale channel enhancement block and spatial intensify block.

4.4.1. HRVI branch analysis

The role of the HRVI branch is to extract spatial texture features from the HRVI image and inject it into the NDVI branch, thereby enhancing the reconstruction quality of HRNDVI. To validate its effectiveness, we train our NDVI-Net without the HRVI branch. The difference of the results is shown in Fig. 11. It can be clearly seen that the result with the HRVI branch contains richer spatial texture details, which are closer to those in the reference HRNDVI. On the contrary, the result without the HRVI branch suffers from blurred details. As a result, this proves the importance of the NDVI branch in reconstructing the texture.

4.4.2. MS-CE block analysis

The multi-scale channel enhancement block is used to extract and screen important features in two branches in the first stage. Specifically, what we need is the intensity distribution characteristics in LRNDVI and the spatial texture features in HRVI. In order to verify the role of the MS-CE block, we replace it with the ordinary convolutional layer. The results are shown in Fig. 12. When there is no MS-CE block, artificial white shadows appear in the local blocks. Conversely, the result with the MS-CE block contains no such distortion, which is closer to the reference HRNDVI. This experiment shows that the MS-CE block is important in preliminary feature extraction, which can observe the input from different scales and cooperate with the channel attention mechanism to select more favorable features and suppress invalid or negative features, such as white shadows in highlighted regions.

4.4.3. SI block analysis

The small details of NDVI are very important in vegetation detection, because they are usually dividing lines between vegetation, such as roads, buildings and so on. The spatial intensify block is to selectively enhance or suppress all the features of the corresponding spatial

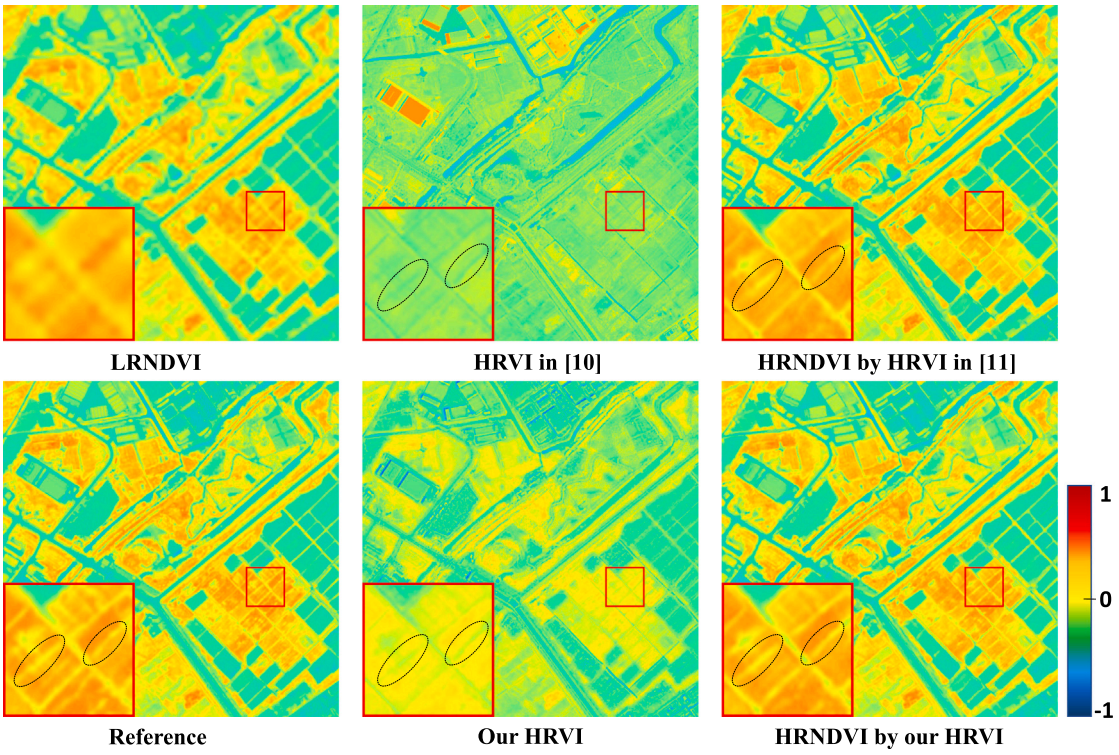

Fig. 10. Validation of HRVI definition. We use the original HRVI (Tu et al., 2009) and our modified HRVI. The corresponding fused results are provided.

Table 1
Quantitative comparison of eight methods on 25 test images from the QuickBird dataset. **Bold** indicates the best result.

Method	RMSE	GMSD	SSIM	CC	VIF	IFC
BDSD (Garzelli et al., 2007)	10.850 ± 3.130	0.115 ± 0.033	0.352 ± 0.147	0.929 ± 0.047	0.245 ± 0.056	1.231 ± 0.472
PRACS (Choi et al., 2010)	7.739 ± 2.047	0.093 ± 0.024	0.342 ± 0.147	0.948 ± 0.042	0.246 ± 0.050	1.355 ± 0.525
PMGI (Zhang et al., 2020a)	8.494 ± 1.847	0.113 ± 0.023	0.267 ± 0.120	0.939 ± 0.043	0.199 ± 0.028	1.106 ± 0.360
MTF-GLP (Aiazzi et al., 2006)	9.931 ± 1.723	0.126 ± 0.027	0.286 ± 0.120	0.925 ± 0.042	0.232 ± 0.042	1.222 ± 0.425
PNN (Masi et al., 2016)	15.492 ± 2.298	0.131 ± 0.017	0.353 ± 0.158	0.905 ± 0.035	0.223 ± 0.050	1.138 ± 0.472
LGC (Fu et al., 2019)	6.238 ± 2.319	0.064 ± 0.027	0.474 ± 0.173	0.964 ± 0.035	0.385 ± 0.078	2.243 ± 0.811
NTV (Zhang et al., 2020b)	6.528 ± 1.311	0.083 ± 0.030	0.403 ± 0.154	0.965 ± 0.022	0.301 ± 0.052	1.648 ± 0.601
NDVI-Net	**4.724 ± 1.048**	**0.043 ± 0.014**	**0.513 ± 0.210**	**0.982 ± 0.012**	**0.417 ± 0.088**	**2.303 ± 0.890**

Table 2
Quantitative comparison of eight methods on 25 test images from the GF-2 dataset. **Bold** indicates the best result.

Method	RMSE	GMSD	SSIM	CC	VIF	IFC
BDSD (Garzelli et al., 2007)	11.047 ± 2.199	0.118 ± 0.024	0.456 ± 0.087	0.908 ± 0.032	0.278 ± 0.043	1.548 ± 0.153
PRACS (Choi et al., 2010)	6.938 ± 0.990	0.087 ± 0.014	0.498 ± 0.096	0.946 ± 0.017	0.307 ± 0.033	1.946 ± 0.313
PMGI (Zhang et al., 2020a)	7.536 ± 1.190	0.105 ± 0.017	0.438 ± 0.092	0.937 ± 0.019	0.272 ± 0.027	1.788 ± 0.166
MTF-GLP (Aiazzi et al., 2006)	7.550 ± 0.967	0.095 ± 0.013	0.465 ± 0.096	0.937 ± 0.015	0.287 ± 0.030	1.816 ± 0.272
PNN (Masi et al., 2016)	12.555 ± 2.957	0.124 ± 0.026	0.500 ± 0.076	0.895 ± 0.052	0.327 ± 0.066	1.768 ± 0.234
LGC (Fu et al., 2019)	6.449 ± 1.524	0.061 ± 0.022	0.592 ± 0.073	0.953 ± 0.024	0.418 ± 0.047	2.738 ± 0.282
NTV (Zhang et al., 2020b)	5.923 ± 1.030	0.064 ± 0.019	0.601 ± 0.082	0.961 ± 0.013	0.392 ± 0.037	2.500 ± 0.273
NDVI-Net	**4.340 ± 0.683**	**0.036 ± 0.010**	**0.740 ± 0.100**	**0.979 ± 0.006**	**0.549 ± 0.034**	**3.450 ± 0.387**

position along the pixel, so it can further strengthen the preservation of small details on the basis of the previous MS-CE block. We conduct an ablation experiment to verify this point, and the results are shown in Fig. 13. Obviously, the result with SI block can better retain those tiny textures, such as gaps and edges between vegetation as highlighted, while the result without SI block cannot.

4.5. Visualization of spatial injection

The spatial texture features in the HRVI branch are injected unidirectionally into the NDVI branch, thereby providing spatial information for HRNDVI reconstruction. In order to show this process intuitively, we randomly select two channels from the feature maps of spatial injection in the HRVI branch and from the features maps before and after the injection in the NDVI branch for visualization, which are shown in Fig. 14.

It can be seen that the features injected into the NDVI branch from the HRVI branch contain very rich spatial texture information. In addition, the output features of the NDVI layer 1 of the NDVI branch have the checkerboard effect caused by transpose convolution, and the texture details are blurred. With the input of spatial features from the HRVI branch, the texture details of features in the NDVI branch are gradually enriched, and the checkerboard effect is gradually eliminated. The visualization experiment further proves the important role of the HRVI branch for HRNDVI reconstruction.

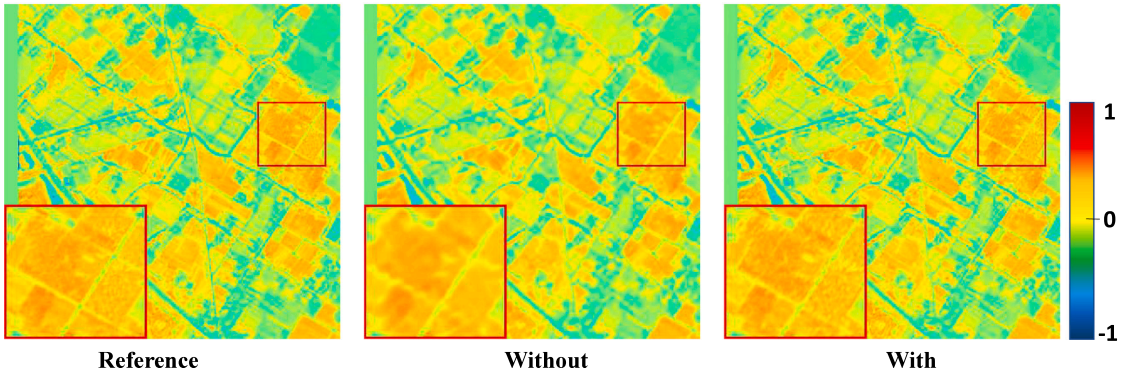

Fig. 11. Ablation experiment of HRVI branch. From left to right: reference NDVI, the result without HRVI branch and the result with HRVI branch.

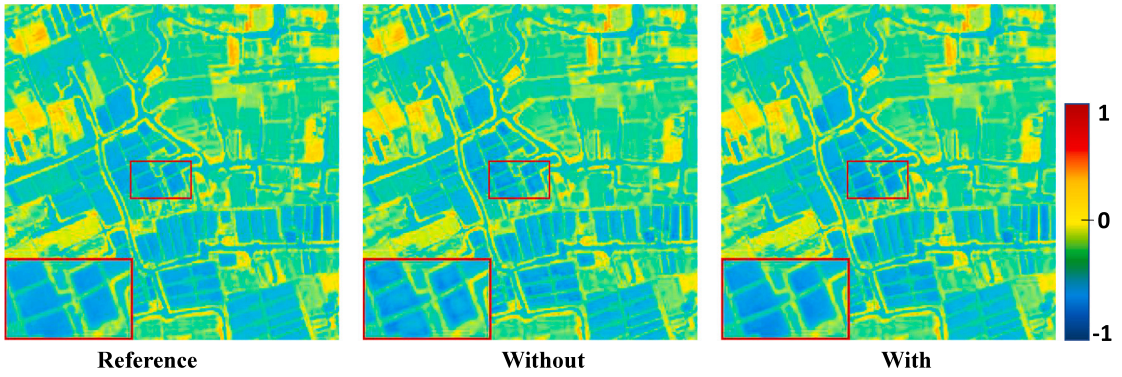

Fig. 12. Ablation experiment of MS-CE block. From left to right: reference NDVI, the result without MS-CE block and the result with MS-CE block.

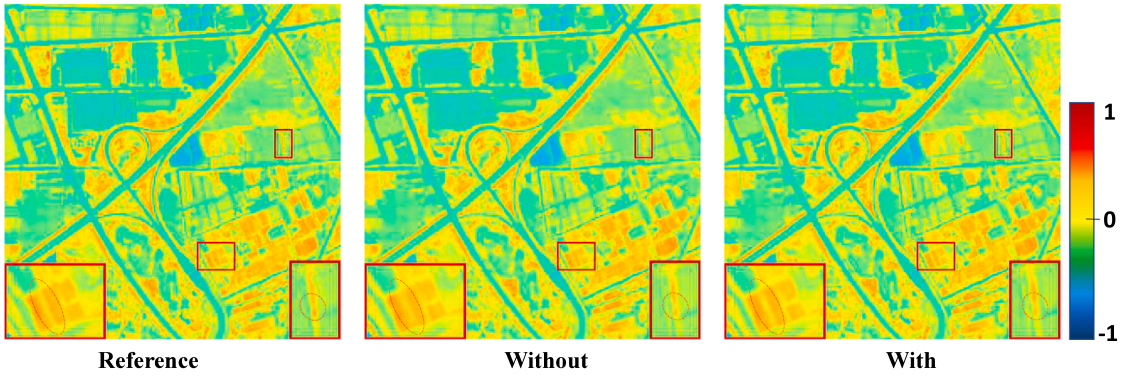

Fig. 13. Ablation experiment of SI block. From left to right: reference NDVI, the result without SI block and the result with SI block.

4.6. Generalization experiment

The generalization ability of deep learning-based methods is an important basis for measuring the performance of a method. In the field of remote sensing image fusion, the model trained on one dataset is difficult to transfer to another due to the different imaging sensors mounted on different satellites. In order to verify the generalization performance of our NDVI-Net, we train it on the GF-2 dataset and then test it on the QuickBird dataset. PNN is also processed in the same way. The experimental results are shown in Fig. 15 and Table 3. It is worth noting that the "Normal" in Fig. 15 and Table 3 means the training and testing of the network are both implemented on the QuickBird dataset.

As can be seen from Fig. 15, the test result of transferred PNN on the QuickBird dataset has a large intensity distortion, and its style is more similar to the GF-2 data in Fig. 8. In contrast, although our method also suffers from performance degradation after transferring, it can still get a relatively good result. The objective indicators in Table 3 are consistent with the qualitative results. Compared with PNN, our NDVI-Net has less performance degradation.

4.7. Application to vegetation detection and enhancement

The NDVI is widely used to analyze the growth status of vegetation. We apply the NDVI to vegetation detection and enhancement. The high-precision HRNDVI generated by our NDVI-Net can overcome the poor accuracy of vegetation detection and enhancement in the HRMS image.

4.7.1. Application strategy

We first give a specific application strategy of NDVI in vegetation detection and enhancement. The detailed description of the detection and enhancement method is shown in Fig. 16.

The value of NDVI is in the range of $[-1, 1]$, which indicates the status of land covered by vegetation. Therefore, in order to detect whether a certain pixel position is covered by vegetation, a threshold

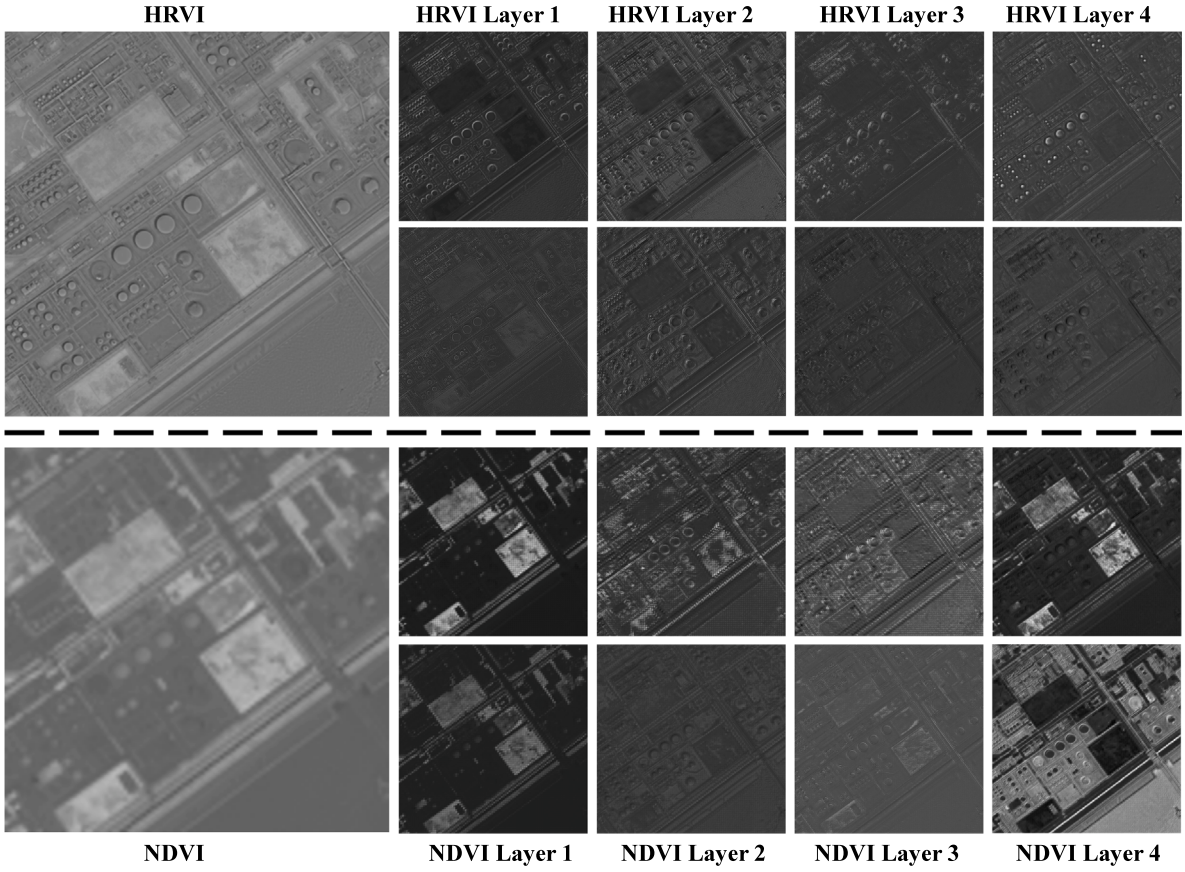

Fig. 14. Visualization of spatial injection.

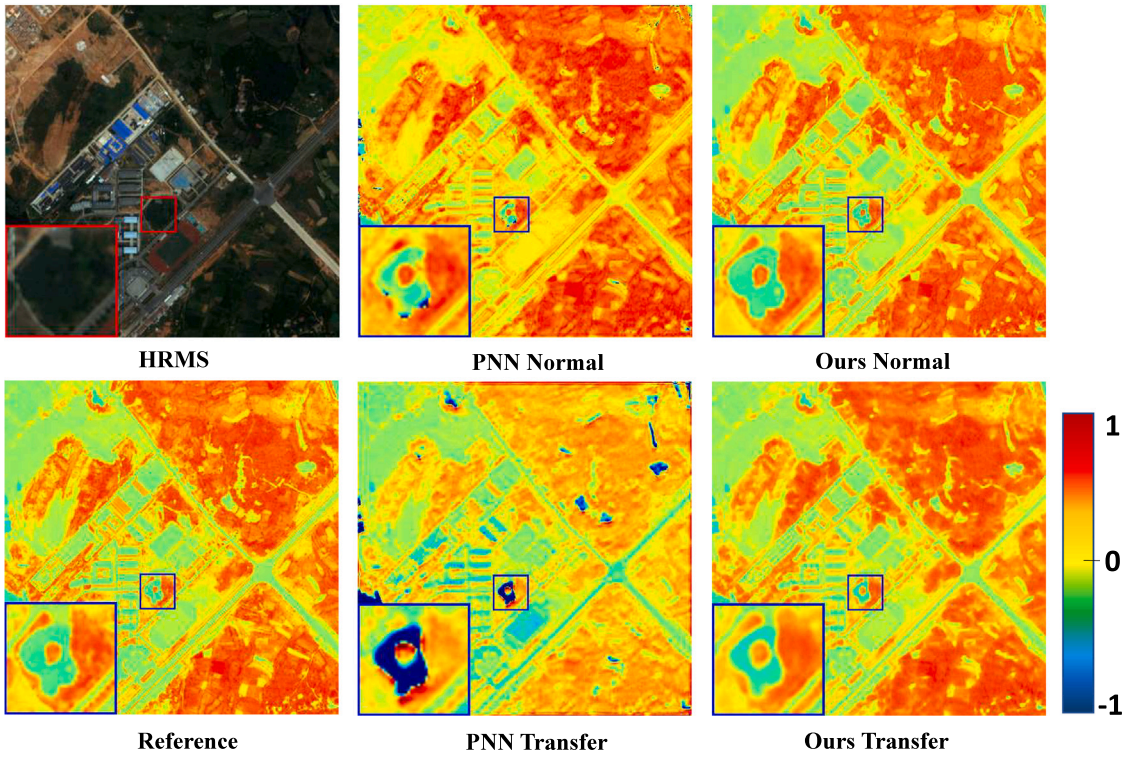

Fig. 15. Visualization of generalization experiment.

Table 3
Quantitative comparison of generalization on 25 test images from the QuickBird dataset.

Method	RMSE	GMSD	SSIM	CC	VIF	IFC
PNN normal (Masi et al., 2016)	15.492 ± 2.298	0.131 ± 0.017	0.353 ± 0.158	0.905 ± 0.035	0.223 ± 0.050	1.138 ± 0.472
PNN transfer (Masi et al., 2016)	56.910 ± 17.266	0.281 ± 0.026	0.014 ± 0.066	0.031 ± 0.203	0.030 ± 0.033	0.127 ± 0.172
PNN degradation	41.418	0.150	0.339	0.874	0.193	1.011
Ours normal	4.724 ± 1.048	0.043 ± 0.014	0.513 ± 0.210	0.982 ± 0.012	0.417 ± 0.088	2.303 ± 0.890
Ours transfer	6.577 ± 1.390	0.074 ± 0.020	0.389 ± 0.146	0.965 ± 0.022	0.289 ± 0.049	1.519 ± 0.518
Ours degradation	1.853	0.031	0.124	0.017	0.128	0.784

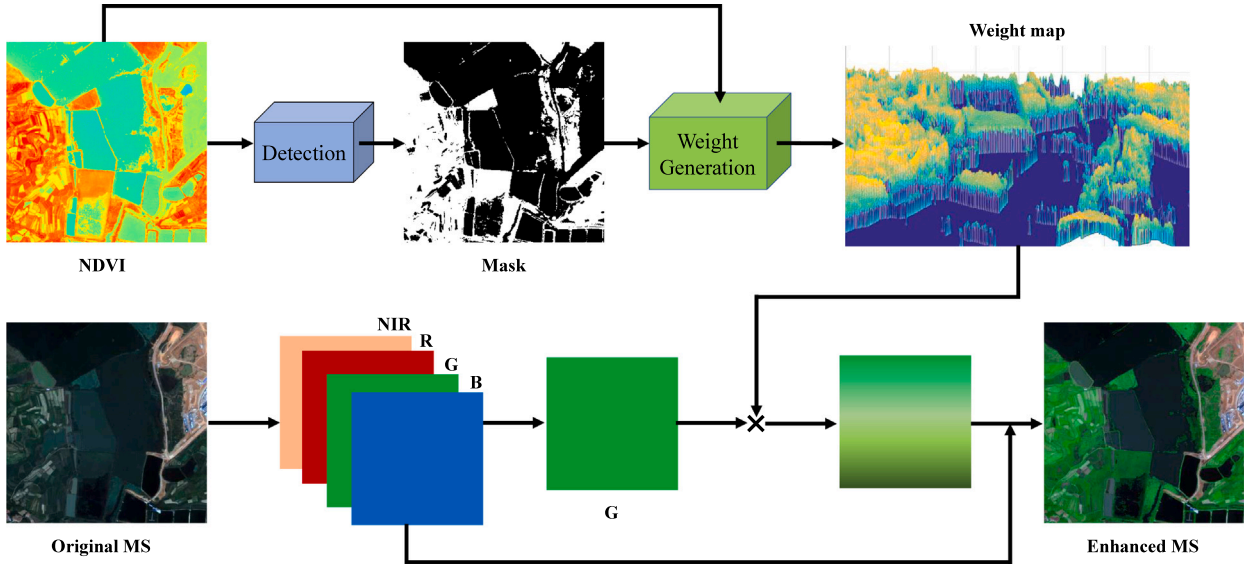

Fig. 16. Schematic diagram of vegetation detection and enhancement. We use HRNDVI generated by NDVI-Net to detect and enhance vegetation.

can be manually set for NDVI as the dividing line between vegetation presence and absence. Therefore, the vegetation detection mask can be obtained by the following rule:

$$Mask_{i,j} = \begin{cases} 1, & NDVI_{i,j} \geq \psi, \\ 0, & \text{others}, \end{cases} \tag{8}$$

where ψ is the threshold, which is a constant. The value of NDVI is not only related to the characteristics of the remote sensing imaging regions, but also related to the response characteristics of sensors. Therefore, the threshold ψ can be adjusted according to the actual situation.

Vegetation enhancement needs to be considered from two aspects. The first is which regions should be enhanced, and the second is what rules should be used for this enhancement. For the first consideration, we should enhance the areas covered by vegetation, while for other areas their original values should be kept to unchanged. For the second consideration, this enhancement should be gradual, that is, the texture structure between vegetation should be strictly maintained. Taking together these two considerations, the nonlinear weight map based on the value of NDVI itself can be generated according to:

$$Weight_{i,j} = e^{\beta \cdot NDVI_{i,j} \cdot Mask_{i,j}}, \tag{9}$$

where β is a scaling factor, which can control the degree of enhancement. It is worth noting that β should be controlled not to exceed the upper numerical limit to avoid truncation errors, and hence it can be adjusted according to actual needs. The nonlinear weight map based on NDVI can effectively maintain the texture between vegetation.

Subsequently, the obtained weight map can be used to enhance the green band of the original MS image, which is formalized as:

$$G_{\text{enhanced}} = G_{\text{original}} \cdot Weight, \tag{10}$$

in which G_{original} is the green band of the original MS image, and G_{enhanced} is enhanced green band. Finally, the enhanced MS image is generated by concatenating the enhanced green band and original blue, red and near infrared bands.

4.7.2. Results analysis

The vegetation detection and enhancement results based on NDVI of different methods are provided in Figs. 17 and 18. It is worth noting that the ψ and β in Eqs. (8) and (9) are controlled unchanged in all methods. In particular, we set $\psi = 0.3$ and $\beta = 0.7$.

Among the vegetation detection masks of all methods, the mask obtained by our method is more consistent with the reference mask. For example, in the highlighted region in Fig. 17, our method can detect the green belt on the highway and the gap between them more accurately. In contrast, other methods are either unable to detect them, or the detected texture is not fine enough. Specifically, BDSD, PRACS, PMGI, MTF-GLP and NTV are failed to detect the green belt, while the green belts detected by PNN and LGC are too thick or incomplete. In addition, the villages and small roads detected by our method are clear and regular, which are consistent with the reference. The results of vegetation enhancement in Fig. 18 are generated based on the NDVI and the vegetation detection results in Fig. 17. It can be seen that in the results of our method, the transition between the enhanced vegetation and the villages is very natural. And the green belt on the highway is also enhanced from invisible to obvious. In contrast, due to the relatively large distortion of NDVI, other methods have led to misjudgments in subsequent vegetation detection and enhancement. In general, compared with other methods, our enhanced result is the closest to the reference enhanced result. We also provide quantitative results in Table 4. It can be seen that our method can obtain the best objective metrics both in vegetation detection and vegetation enhancement.

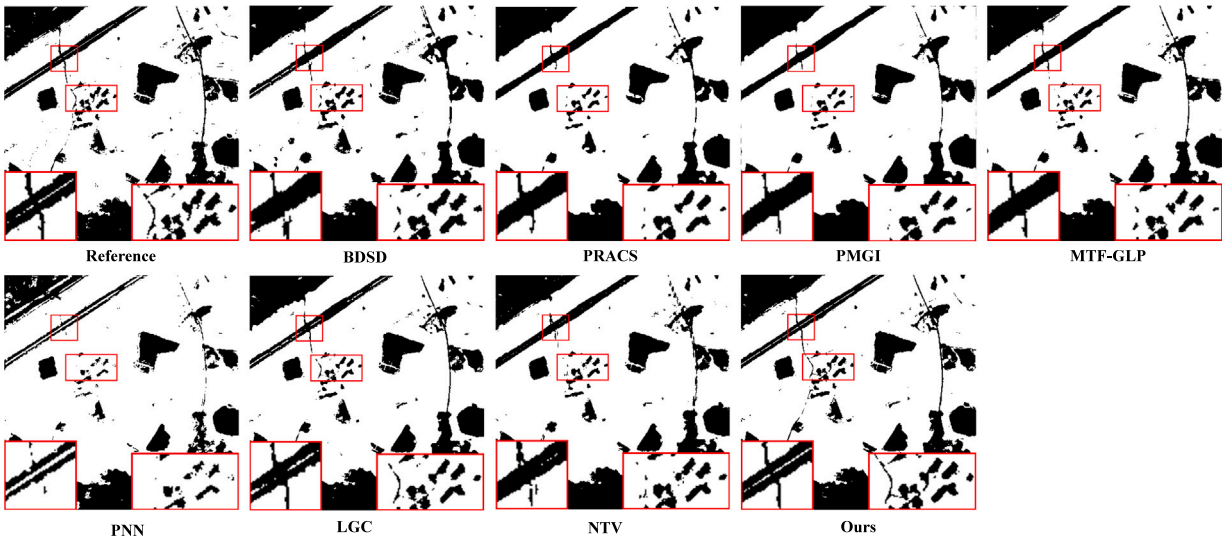

Fig. 17. Qualitative results of vegetation detection. The images are detection masks by reference, BDSD (Garzelli et al., 2007), PRACS (Choi et al., 2010), PMGI (Zhang et al., 2020a), MTF-GLP (Aiazzi et al., 2006), PNN (Masi et al., 2016), LGC (Fu et al., 2019), NTV (Zhang et al., 2020b) and our NDVI-Net.

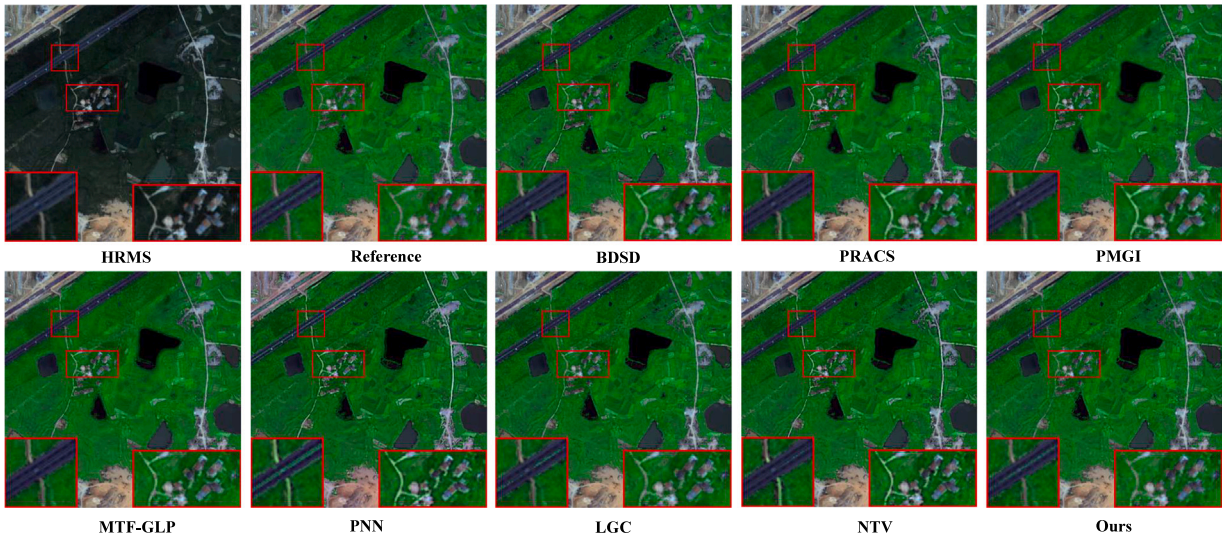

Fig. 18. Qualitative results of vegetation enhancement. The images are HRMS image, enhanced results by reference, BDSD (Garzelli et al., 2007), PRACS (Choi et al., 2010), PMGI (Zhang et al., 2020a), MTF-GLP (Aiazzi et al., 2006), PNN (Masi et al., 2016), LGC (Fu et al., 2019), NTV (Zhang et al., 2020b) and our NDVI-Net. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Quantitative comparison of the vegetation detection and enhancement. **Bold** indicates the best result.

Method	Detection	Enhancement					
	RMSE	RMSE	GMSD	SSIM	CC	VIF	IFC
BDSD (Garzelli et al., 2007)	0.225 ± 0.061	4.786 ± 1.056	0.062 ± 0.007	0.627 ± 0.05	0.908 ± 0.024	0.504 ± 0.104	4.365 ± 0.235
PRACS (Choi et al., 2010)	0.223 ± 0.067	4.324 ± 1.140	0.055 ± 0.010	0.625 ± 0.055	0.918 ± 0.029	0.464 ± 0.082	4.199 ± 0.169
PMGI (Zhang et al., 2020a)	0.237 ± 0.070	4.735 ± 1.177	0.064 ± 0.012	0.533 ± 0.065	0.902 ± 0.032	0.409 ± 0.082	3.857 ± 0.132
MTF-GLP (Aiazzi et al., 2006)	0.226 ± 0.067	4.779 ± 0.964	0.060 ± 0.009	0.415 ± 0.017	0.899 ± 0.023	0.382 ± 0.037	3.673 ± 0.153
PNN (Masi et al., 2016)	0.321 ± 0.065	7.023 ± 1.115	0.090 ± 0.002	0.643 ± 0.088	0.844 ± 0.025	0.454 ± 0.123	3.273 ± 0.009
LGC (Fu et al., 2019)	0.194 ± 0.062	3.702 ± 1.051	0.037 ± 0.011	0.737 ± 0.061	0.940 ± 0.023	0.612 ± 0.107	5.278 ± 0.435
NTV (Zhang et al., 2020b)	0.224 ± 0.059	4.409 ± 0.920	0.056 ± 0.001	0.743 ± 0.068	0.919 ± 0.021	0.593 ± 0.120	4.969 ± 0.414
NDVI-Net	**0.176 ± 0.055**	**3.264 ± 0.936**	**0.029 ± 0.009**	**0.797 ± 0.047**	**0.953 ± 0.018**	**0.671 ± 0.101**	**5.467 ± 0.278**

4.8. Application to land cover mapping

NDVI can also characterize some other types of land cover. Generally, negative values indicate that the ground is covered by water and snow; 0 represents that there is rock or bare soil; positive values refer to that there is vegetation, which increases with increasing coverage. Therefore, we can implement land cover mapping based on NDVI, in which the coverings are divided into three categories, saying water, bare land or buildings, and vegetation. The land cover mapping results based on NDVI of different methods are provided in Fig. 19. It can be seen that our results are the most consistent with the reference, which can more accurately classify the three types of coverings. We also provide quantitative results for a more comprehensive comparison, as

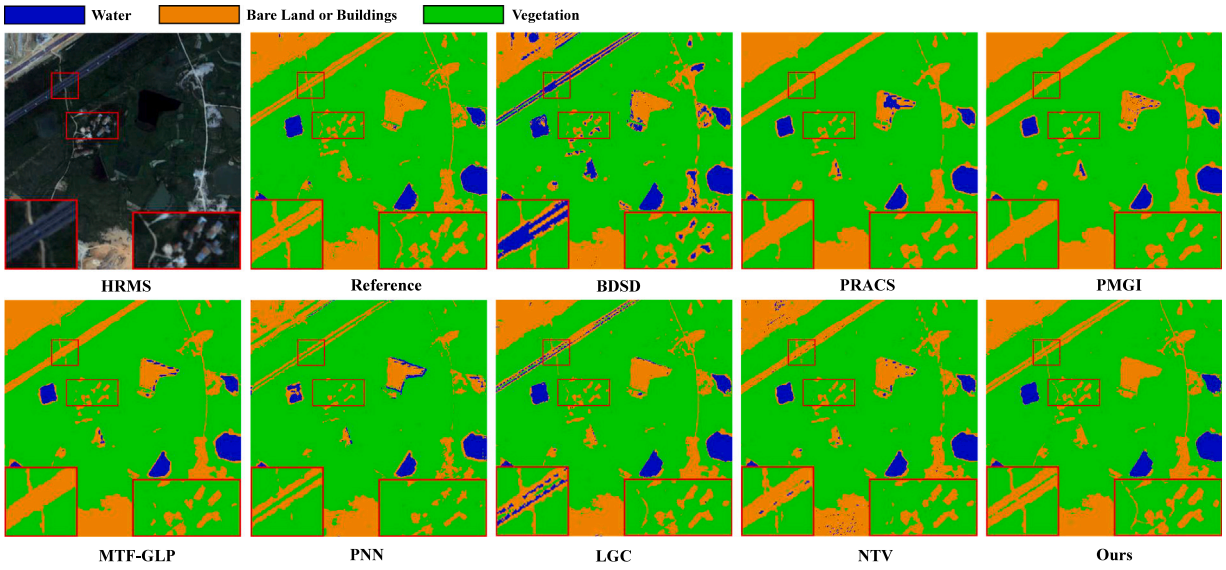

Fig. 19. Qualitative results of land cover mapping. The images are the HRMS, the results by reference, BDSD (Garzelli et al., 2007), PRACS (Choi et al., 2010), PMGI (Zhang et al., 2020a), MTF-GLP (Aiazzi et al., 2006), PNN (Masi et al., 2016), LGC (Fu et al., 2019), NTV (Zhang et al., 2020b) and our NDVI-Net.

Table 5
Quantitative comparison of the land cover mapping. **Bold** indicates the best result.

Method	RMSE	GMSD	SSIM	CC	VIF	IFC
BDSD (Garzelli et al., 2007)	24.381 ± 2.237	0.097 ± 0.005	0.728 ± 0.024	0.834 ± 0.014	0.568 ± 0.040	0.882 ± 0.268
PRACS (Choi et al., 2010)	20.734 ± 2.944	0.096 ± 0.008	0.752 ± 0.040	0.884 ± 0.008	0.515 ± 0.017	0.873 ± 0.205
PMGI (Zhang et al., 2020a)	24.160 ± 5.135	0.109 ± 0.006	0.691 ± 0.069	0.846 ± 0.039	0.406 ± 0.032	0.505 ± 0.057
MTF-GLP (Aiazzi et al., 2006)	20.791 ± 3.482	0.095 ± 0.010	0.752 ± 0.050	0.885 ± 0.019	0.530 ± 0.023	0.906 ± 0.177
PNN (Masi et al., 2016)	29.209 ± 3.079	0.116 ± 0.008	0.663 ± 0.047	0.767 ± 0.007	0.372 ± 0.043	0.441 ± 0.071
LGC (Fu et al., 2019)	17.724 ± 2.010	0.075 ± 0.004	0.812 ± 0.016	0.915 ± 0.006	**0.605 ± 0.012**	1.315 ± 0.285
NTV (Zhang et al., 2020b)	20.656 ± 2.734	0.091 ± 0.007	0.746 ± 0.026	0.888 ± 0.013	0.545 ± 0.006	0.922 ± 0.150
NDVI-Net	**16.973 ± 3.339**	**0.074 ± 0.007**	**0.828 ± 0.035**	**0.924 ± 0.018**	0.604 ± 0.016	**1.356 ± 0.149**

shown in Table 5. Our method achieved the best results on five metrics, which further proves the good performance of our method.

5. Conclusion

In this paper, a novel two-branch NDVI fusion network based on multi-scale and attention mechanism, called NDVI-Net, is proposed to generate the high-resolution NDVI with accurate intensity distribution and clear texture details. The HRVI is introduced to our model to provide spatial texture information for reconstructing HRNDVI. We first adopt the multi-scale channel enhancement blocks to extract and screen features separately from the NDVI branch and HRVI branch. Meanwhile, the texture detail features in the HRVI branch are unidirectionally injected into the NDVI branch to provide spatial information for HRNDVI reconstruction. Subsequently, the spatial intensify blocks are used for pixel-by-pixel feature selection along spatial locations, which can further enhance the retention of minute details. Under the constraints of a specific loss function, the high-quality HRNDVI can be obtained. Extensive qualitative and quantitative experiments demonstrate the advantages of our NDVI-Net over state-of-the-art methods in terms of both subjective visual effect and quantitative metrics. Moreover, our NDVI-Net has good generalization performance, which can be better migrated to other satellite data. The expanded application to vegetation detection and enhancement, and land cover mapping further proves the advantages of our method.

Our NDVI-Net needs the reference HRNDVI during the training phase to guide the optimization of the network, which may limit its use on some datasets. For example, in those datasets with a small number and small image size, the training data obtained by down-sampling cannot train the network adequately. In the future, we will focus on the research of unsupervised fusion network for generating HRNDVI, and apply it to a wider range of remote sensing tasks, such as grain yield prediction, land-cover change detection and crop identification.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was sponsored in part by the National Natural Science Foundation of China (61773295 and 61971315), and in part by the Natural Science Foundation of Hubei Province, China (2018CFB435 and 2019CFA037).

References

Aiazzi, B., Alparone, L., Baronti, S., Garzelli, A., Selva, M., 2006. Mtf-tailored multiscale fusion of high-resolution ms and pan imagery. Photogramm. Eng. Remote Sens. 72 (5), 591–596.

Aiazzi, B., Baronti, S., Selva, M., 2007. Improving component substitution pansharpening through multivariate regression of ms + pan data. IEEE Trans. Geosci. Remote Sens. 45 (10), 3230–3239.

Aiazzi, B., Baronti, S., Selva, M., Alparone, L., 2013. Bi-cubic interpolation for shift-free pan-sharpening. ISPRS J. Photogramm. Remote Sens. 86, 65–76.

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Carlson, T.N., Ripley, D.A., 1997. On the relation between ndvi, fractional vegetation cover, and leaf area index. Remote Sens. Environ. 62 (3), 241–252.

Carper, W., Lillesand, T., Kiefer, R., 1990. The use of intensity-hue-saturation transformations for merging spot panchromatic and multispectral image data. Photogramm. Eng. Remote Sens. 56 (4), 459–467.

Chen, C., Li, Y., Liu, W., Huang, J., 2015. Sirf: Simultaneous satellite image registration and fusion in a unified framework. IEEE Trans. Image Process. 24 (11), 4213–4224.
Choi, J., Yu, K., Kim, Y., 2010. A new adaptive component-substitution-based satellite image fusion by using partial replacement. IEEE Trans. Geosci. Remote Sens. 49 (1), 295–309.
Deshmukh, M., Bhosale, U., 2010. Image fusion and image quality assessment of fused images. Int. J. Image Process. 4 (5), 484.
Duran, J., Buades, A., Coll, B., Sbert, C., Blanchet, G., 2017. A survey of pansharpening methods with a new band-decoupled variational model. ISPRS J. Photogramm. Remote Sens. 125, 78–105.
Fu, X., Lin, Z., Huang, Y., Ding, X., 2019. A variational pan-sharpening with local gradient constraints. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10265–10274.
Garzelli, A., Nencini, F., Capobianco, L., 2007. Optimal mmse pan sharpening of very high resolution multispectral images. IEEE Trans. Geosci. Remote Sens. 46 (1), 228–236.
Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141.
Hua, Y., Mou, L., Zhu, X.X., 2019. Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification. ISPRS J. Photogramm. Remote Sens. 149, 188–199.
Itti, L., Koch, C., Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. (11), 1254–1259.
Johnson, B., 2014. Effects of pansharpening on vegetation indices. ISPRS Int. J. Geo-Inf. 3 (2), 507–522.
Liu, X., Wang, Y., Liu, Q., 2018. Psgan: a generative adversarial network for remote sensing image pan-sharpening. In: Proceedings of the IEEE International Conference on Image Processing. pp. 873–877.
Luong, M.-T., Pham, H., Manning, C.D., 2015. Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025.
Ma, J., Xu, H., Jiang, J., Mei, X., Zhang, X.-P., 2020a. Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion. IEEE Trans. Image Process. 29, 4980–4995.
Ma, J., Yu, W., Chen, C., Liang, P., Guo, X., Jiang, J., 2020b. Pan-gan: An unsupervised pan-sharpening method for remote sensing image fusion. Inf. Fusion 62, 110–120.
Ma, J., Yu, W., Liang, P., Li, C., Jiang, J., 2019. Fusiongan: A generative adversarial network for infrared and visible image fusion. Inf. Fusion 48, 11–26.
Ma, J., Zhang, H., Yi, P., Wang, Z., 2020c. Scscn: A separated channel-spatial convolution net with attention for single-view reconstruction. IEEE Trans. Ind. Electron. 67 (10), 8649–8658.
Masi, G., Cozzolino, D., Verdoliva, L., Scarpa, G., 2016. Pansharpening by convolutional neural networks. Remote Sens. 8 (7), 594.
Rahmani, S., Strait, M., Merkurjev, D., Moeller, M., Wittman, T., 2010. An adaptive ihs pan-sharpening method. IEEE Geosci. Remote Sens. Lett. 7 (4), 746–750.
Sheikh, H.R., Bovik, A.C., 2006. Image information and visual quality. IEEE Trans. Image Process. 15 (2), 430–444.
Sheikh, H.R., Bovik, A.C., De Veciana, G., 2005. An information fidelity criterion for image quality assessment using natural scene statistics. IEEE Trans. Image Process. 14 (12), 2117–2128.
Tian, X., Chen, Y., Yang, C., Gao, X., Ma, J., 2020. A variational pansharpening method based on gradient sparse representation. IEEE Signal Process. Lett. 27, 1180–1184.
Tu, T.-M., Lu, H.-T., Chang, Y.-C., Chang, J.-C., Chang, C.-P., 2009. A new vegetation enhancement/extraction technique for ikonos and quickbird imagery. IEEE Geosci. Remote Sens. Lett. 6 (2), 349–353.
Tucker, C.J., 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sens. Environ. 8, 127–150.
Wald, L., Ranchin, T., Mangolini, M., 1997. Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images. Photogramm. Eng. Remote Sens. 63 (6), 691–699.
Wang, Z., Bovik, A.C., 2002. A universal image quality index. IEEE Signal Process. Lett. 9 (3), 81–84.
Wang, Q., Shi, W., Atkinson, P.M., 2016. Area-to-point regression kriging for pan-sharpening. ISPRS J. Photogramm. Remote Sens. 114, 151–165.
Woo, S., Park, J., Lee, J.-Y., So Kweon, I., 2018. Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision. pp. 3–19.
Xu, H., Ma, J., Zhang, X.-P., 2020. Mef-gan: Multi-exposure image fusion via generative adversarial networks. IEEE Trans. Image Process. 29, 7203–7216.
Xue, W., Zhang, L., Mou, X., Bovik, A.C., 2013. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. IEEE Trans. Image Process. 23 (2), 684–695.
Yang, F., Matsushita, B., Fukushima, T., Yang, W., 2012. Temporal mixture analysis for estimating impervious surface area from multi-temporal modis ndvi data in japan. ISPRS J. Photogramm. Remote Sens. 72, 90–98.
Zhang, H., Xu, H., Xiao, Y., Guo, X., Ma, J., 2020. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 12797–12804.
Zhang, M., Zhao, Z., Chen, Y., Wang, Z., Tian, X., 2020. Fusionndvi: A novel fusion method for ndvi in remote sensing. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 4816–4820.
Zhao, H., Gallo, O., Frosio, I., Kautz, J., 2016. Loss functions for image restoration with neural networks. IEEE Trans. Comput. Imaging 3 (1), 47–57.
Zhu, X., Liu, D., 2015. Improving forest aboveground biomass estimation using seasonal landsat ndvi time-series. ISPRS J. Photogramm. Remote Sens. 102, 222–231.