Journal of Network Intelligence ©2021 ISSN 2414-8105 (Online)
Taiwan Ubiquitous Information Volume **6**, Number **1**, February **2021**

Sparse Representation and SRCNN based Spatio-temporal Information Fusion Method of Multi-sensor Remote Sensing Data

Shuai Yang

College of Electrics Engineering
Heilongjiang University
Harbin 150080, China
pengzaiqingdao@163.com

Xiaofei Wang *

College of Electrics Engineering
Heilongjiang University
Harbin 150080, China
*Corresponding author: nk_wxf@hlju.edu.cn

Received March 2019; revised February 2020

Abstract. *The classical spatio-temporal fusion algorithms STARFM and SPSTFM will have large fusion errors when the phenological changes or type changes appear. In this paper, based on the spatial feature information of the image, we proposed a new spatio-temporal information fusion method which combines SRCNN (Super-Resolution Convolutional Neural Network) and sparse representation. Firstly, complete the feature reconstruction of the reflectance change image by combining SRCNN and sparse representation, and then the reconstructed image is superimposed by the time weight to obtain the predicated reflectance image. Experiments show that the proposed method is better than the classic spatio-temporal fusion algorithms STARFM and SPSTFM.*

Keywords: Spatio-temporal fusion, machine learning, SRCNN, sparse representation.

1. **Introduction.** With the continuous development of remote sensing technology, remote sensing technology can provide various information about crop ecological environment and crop growth objectively, accurately, and timely. It is an important source for accurate field data. However, under the constraints of the hardware technical conditions of existing satellite sensors and the cost of satellite launching, remote sensing satellites cannot obtain remote sensing images with multi-attribute fine resolution, which restricts the application of remote sensing images. For example, Landsat satellites obtain images with spatial resolutions in the 30m range, whereas a 16-day return visit cycle limits its use to detect rapid land changes, on the other hand, medium resolution imaging spectra carried on Terra/Aqua satellites radiometer (MODIS) provides daily observations, but having a coarse spatial resolution of 250-1000m is not conducive to monitoring land cover changes in heterogeneous landscapes. Therefore, spatio-temporal fusion has emerged as a method of providing satellite imagery with fine temporal and spatial resolution.

During recent decades, the method of spatio-temporal fusion has been widely applied and mentioned. Gao et al. (2006) first proposed the Spatial And Temporal Adaptive Reflectance Fusion Model (STARFM) [1] for the identification of surface cover types of fractured patches, which can effectively eliminate singular points and is suitable for detecting the gradual changes of the large-scale range of space. Zhu et al. (2010) proposed Enhanced Spatial and Temporal Adaptive Reflectance Fusion Model (ESTARFM) [2] based on STARFM, which is more accurate and efficient for complex and heterogeneous features, and it solved the "time smoothing" problem of the STARFM algorithm. Both STARFM and ESTARFM are weight-based spatio-temporal fusion algorithms, which are limited to fine-grained landscapes and will reduce the accuracy of their fused images in fine-grained heterogeneous landscapes. Hilker et al. (2009) proposed a Spatial-Temporal Adaptive Algorithm for mapping Reflectance Change (STAARCH) [3] to observe changes in forest vegetation. Although this algorithm can analyze the change of reflectivity of input images and handle the dynamic change of land cover type, it requires two landscape Landsat images not suitable for areas where image acquisition is difficult. Wu et al. (2012) proposed a Spatial And Temporal Data Fusion Approach (STDFA) [4] based on the time-varying feature of the feature for the extraction of rice area. This algorithm also requires multi-phase image support, which limits the application of the algorithm.

In recent years, machine learning has been widely used in various fields such as human motion recognition [5] and video salient region detection [6], it also can be used to solve problems in the field of remote sensing. and it can be used to solve problems in the field of remote sensing. The spatio-temporal fusion algorithm based on machine learning has received extensive attention in the field of image processing. Hong et al. (2012) proposed a Sparse Representation-based Spatio-temporal Reflectance Fusion Model (SPSTFM) [7]. The model uses the difference images of the MODIS image pairs and Landsat image pairs which on the front and back phases of the predicted phase to train fine-resolution and low-resolution dictionaries representing time-varying features, and then uses the MODIS image of the predicted time to generate a Landsat-like fusion image; Subsequently, Song and Huang (2013) proposed a sparse representation spatio-temporal reflectance fusion model using only one pair of known fine and coarse spatial resolution image [8]. The model first uses the sparse representation algorithm to enhance the MODIS image to obtain an intermediate transition image and then uses a high-pass filtering model to fuse the observed Landsat image and the transition image. This model reduces the number of known image pairs that need to be input, so that the algorithm can be applied in the absence of data, and has universal applicability. The feature-based spatio-temporal fusion methods take into account the spatial information characteristics of the changing image and can make more accurate predictions of the complex surface reflectance image compared to the weight-based methods.

Nowadays, deep learning, as a branch of machine learning, is gradually maturing and it can be applied to medical diagnosis [9] and drone surveillance [10]. Convolutional neural network (CNN) [11] is one of the most representative neural networks in deep learning technology. Convolutional neural network (CNN) can be used to solve problems in computer vision such as object recognition [12]and image classification [13]because of its simple model structure. In addition, CNN also can be used to encoding [14] and decoding. Girshick et al. (2014) proposed region-based Convolutional Neural Networks(R-CNN) [15] which improves mean average precision (MAP). Dong et al. (2016) analogous to sparse coding, proposed Super-Resolution Convolutional Neural Network (SRCNN) [16] model to hide the automatic learning with layers replaces the dictionary modeling operation, cleverly establishing CNN end-to-end mapping between low-resolution and fine resolution images, which greatly improves the reconstruction accuracy and running speed. SRCNN network has the advantages of simple data preprocessing, the convolution learning algorithm is not limited by image patch size, and directly obtains the mapping relationship between fine and coarse images, and can learn from the neighborhood relationship between image patchs to complete the partial contour of the image features. Wang et al. (2020)

proposed a deep learning method [17] based on the technology of Internet of Things and fast R-CNN. Tseng et al. (2020) proposed a new convolutional neural network architecture named DNetUnet [18], which combines U-Nets with different down-sampling levels and a new dense block as feature extractor. In addition, DNetUnet is a semi-supervised learning method, which can be used not only to obtain expert knowledge from the labelled corpus, but also to enhance the performance of learning algorithm generalization ability from unlabelled data.

In view of these trends, in order to reduce the fusion error and achieve better spatial detail reconstruction, this paper improves the SRCNN to make it suitable for remote sensing image super-resolution of small sample training sets. Combining SRCNN theory with SPSTFM theory, an enhanced MODIS and Landsat reflectivity image fusion method based on SPSTFM is proposed. The SRCNN theory is used to train the differential image patches of Landsat and MODIS images to effectively extract the edge structure of the image. The reconstructed images of the relative prediction phases at different observations are superimposed by time weights to achieve the reflectivity fusion of the predicted phase. Finally, the prediction results are compared with actual images, STARFM algorithm prediction results and SPSTFM algorithm prediction results to verify the effectiveness of the algorithm and the improvement of the fusion accuracy.

The remainder of this paper is organized as follows. In Section 2, sparse representation and sparse coding are briefly introduced at first, and then the spatio-temporal fusion algorithm based on sparse representation is introduced in detail. In Section 3, the theory of SRCNN and the proposed method of combining SRCNN and SPSTFM are introduced in detail. The experimental results with actual data are shown in Section 4 which also includes comparisons with STARFM and SPSTFM. Finally, this paper is concluded in Section 5.

The reminder of this paper is organized as follows. In Section 2, we introduce the proposed method, which consists of the learning of the complete grouped basis, the acquisition of stego images and the extraction of secret messages. The results of the experiments and analysis are shown in Section 3. Section 4 concludes our work and outlines our future research direction.

2. **Spatio-temporal fusion based on sparse representation.**

2.1. **Sparse representation.**

Knowing matrix $D \in \mathbf{R}^{m \times n}$, each column of the matrix represents a base atom, n represents the dimension of the atom, and m represents the number of base atoms, where $n < m$. In the field of imagery, sparse representation theory holds that image x can be represented linearly by some of the base atoms in D , namely:

$$x = Da \tag{1}$$

Where $a \in \mathbf{R}^m$ is the sparse representation coefficient and D is the overcomplete dictionary of the image. The purpose of sparse representation is to design an efficient dictionary by sparse coding algorithm and find the representation vector with the least non-zero elements so that the combination of the two can express the original signal most closely. The target expression is as follows:

$$\hat{a} = arg\ min\ \|a\|_0 \tag{2}$$

Where $\hat{a}$ is the estimator of a and $\|a\|_0$ is the l_0-norm of a, i.e. the number of non-zero elements in the vector.

2.2. **Spatio-temporal fusion based on sparse representation.** There are Landsat-MODIS image pairs on t_1 and t_3 and MODIS images on t_2 ($t_1 < t_2 < t_3$) as known conditions, Landsat images with a spatial resolution of 30m are used as fine-resolution images, and MODIS images with a spatial resolution of 500m are used as low-resolution images. Interpolate the MODIS image using bilinear interpolation and extend it to the same size as the Landsat image. It is defined that H_i and L_i are Landsat images and MODIS images on t_i ($t = 1, 2, 3$).

First, construct a difference image of the fine-resolution image pair and the low-resolution image pair: the MODIS difference image from t_1 to t_3 (i.e. the low-resolution difference image LRDI) and the corresponding Landsat difference image of the same period (i.e. the fine-resolution difference image HRDI). The difference images (LRDI and HRDI) themselves have high-frequency detail. There is a close relationship among the difference images (i.e. between t_1 and t_2 , t_2 and t_3 , t_1 and t_3). Therefore, the HRDI between t_1 and t_2 and the HRDI between t_2 and t_3 can be predicted by the known HRDI between t_1 and t_3. In particular, our desired differential image patch can be predicted by a linear combination of structural primitives extracted from known HRDI.

We use and represent HRDI and LRDI between t_i and t_j , respectively h_{ij}, and l_{ij} define and as their corresponding image patches. The relationship diagram of these variables is shown in Fig.1. The predicted results of H_2 are as follows:

$$H_2 = W_1 * (H_1 + H_{21}) + W_2 * (H_3 + H_{32}) \tag{3}$$

Where H_{21} and H_{32} are predicted HRDI, W_1 is a weighted parameter for predicting the Landsat reflectance image on t_2 using the Landsat reflectance image on t_1 , W_3 is similar to W_1. For the determination of the weighting parameters, please refer to Section III-D in [5]. To calculate H_{21} and H_{32} , we first train the HRDI and LRDI image patches between t_1 and t_3 to develop dictionary pairs D_1 and D_m , respectively. Based on sparse representation theory, dictionary pairs can use the following optimization equations:

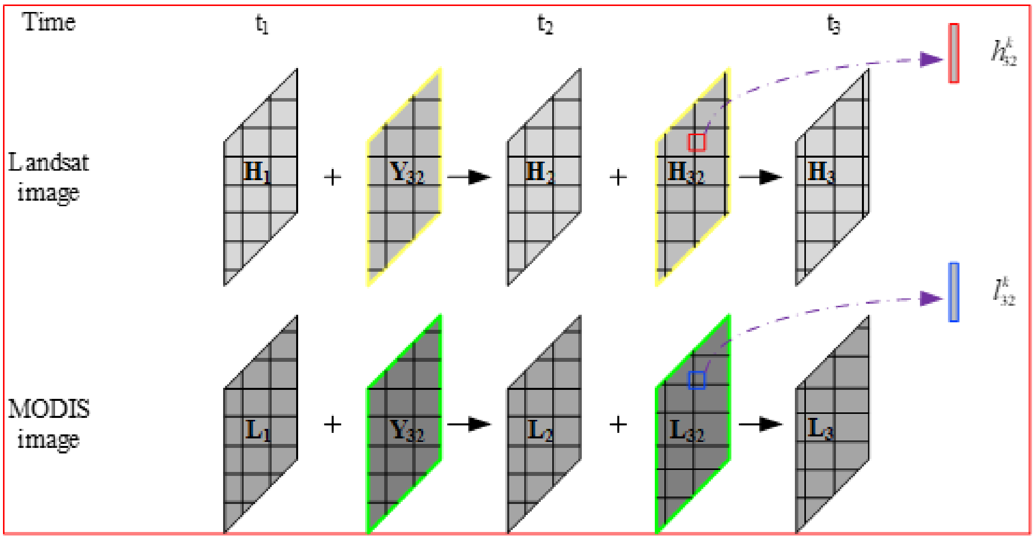


Figure 1. Relationship diagram for the images and difference images and comparison diagram for Landsat image and MODIS image , where h_{32}^k and l_{32}^k stand for the kth patches of H_{32} and L_{32} , respectively.

$$\{D_l^*, D_m^*, a^*\} = \arg\min_{D_l, D_m, a} \{\|H - D_l\Lambda\|_2^2 + \|L - D_m\Lambda\|_2^2 + \lambda\|a\|_1\} \tag{4}$$

Where H and L are column combinations of image patches arranged in lexicographic order from H_{31} and L_{31} , respectively. Similarly, a is a column combination representing the coefficients corresponding to each of H and L. In the algorithm of Yang et al. [19], the training sample is the edge structure of the image, and two dictionaries are learned by optimizing the following function connections:

$$\{D^*, a^*\} = \arg\min_{D,\Lambda}\{\|Z - Da\|_2^2 + \lambda\|a\|_1\} \tag{5}$$

Where $Z = [Y; X]$ and $D = [D_1; D_m]$. Considering the difference in amplitude and variance between HRDI and LRDI, we use the alternate update mode to solve for D_1 , and D_m in (4). For the detailed procedure of the dictionary column update, please refer to [20].

After obtained the dictionaries D_1 and D_m , since the sparse representation coefficients are forced to be considered identical during the dictionary training process, if the sparse representation coefficients of the LRDI image patches are obtained relative to the dictionary D_m , the same sparse representation coefficients and corresponding the dictionary D_1 reconstructs the corresponding HRDI image patch.We use l_{21}^k to represent the kth image patch of L_{21} , and we can estimate the sparse representation coefficient by minimizing the l_1-norm problem:

$$a^* = \arg\min_a f(a) = \arg\min_a \frac{1}{2}\left\|x_{21}^k - D_m a\right\|_2^2 + \lambda\|a\|_1 \tag{6}$$

The corresponding image patch h_{21}^k of HRDI can be expressed as:

$$h_{21}^k = D_1 a^* \tag{7}$$

3. Methodology.

3.1. Theoretical basis of SRCNN algorithm.

Supposing a low-resolution (LR) image was interpolated using the Bicubic interpolation method to obtain the same scale image as a fine-resolution (HR) image. The interpolated image is named Y. The purpose of the algorithm is to use the restored image which is named $F(Y)$ to maximize the proximity to the fine -resolution original image named X. SRCNN algorithm is a 3-layer convolutional network which analogous to the Super-Resolution reconstruction process based on sparse representation. The 3-layer network corresponds to three functions of feature extraction, nonlinear mapping, and image super-resolution. The network structure of the algorithm is shown in Fig.2.

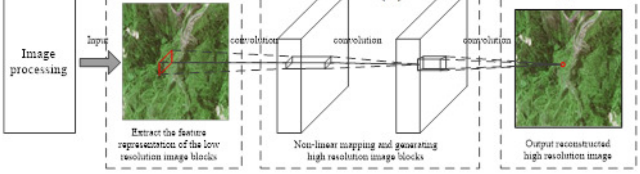


Image processing
Input
Y
convolution
F(Y)
convolution
convolution
X
Extract the feature representation of the low resolution image blocks
Non-linear mapping and generating high resolution image blocks
Output reconstructed high resolution image

Figure 2. Flowchart of SRCNN network structure. The first step is feature extraction, the second step is nonlinear mapping, and the third step is reconstruction.

The first layer of the network extracts and characterizes the image patch, which can be expressed as:

$$F_1(Y) = max(0, W_1 * Y + B_1) \tag{8}$$

Where W_1 and B_1 represent filter and deviation, * represents convolution. The filter named W_1 consists of the number of $c \times f_1 \times f_2$ filters is n_1 ,c is the number of channels included in the image, f_1 is the size of the filter's spatial domain. Therefore, W_1 performed n_1 convolutions on the image, using the $c \times f_1 \times f_1$ as a convolution kernel. This layer outputs five feature maps, B_1 is a n_1-dimensional vector with each element corresponding to a filter. Using ReLU as the activation function, i.e. $max(0, x)$.

The first layer of the network extracts the n_1-dimensional features of the image patch; the second layer of the network maps the n_1-dimensional feature vector to the n_2-dimensional feature vector, which is a feature-to-feature mapping. The second layer of the network can be expressed as:

$$F_2(Y) = max(0, W_2 * F_1(Y) + B_2) \tag{9}$$

Where W_2 contains the number of $n_1 \times f_2 \times f_2$ filters is n_2 , and B_2 is a n_2-dimensional vector.

Traditional methods usually predefine the way of the fusion, such as calculating the average of overlapping part. SRCNN uses a learning method to merge overlapping image patches and complete the process of image reconstruction by using the three-layer network.The third layer of the network can be expressed as:

$$F(Y) = W_3 * F_2(Y) + B_3 \tag{10}$$

Where W_3 contains the number of $n_1 \times f_2 \times f_2$ filters is c , and B_3 is a c-dimensional vector.

According to equations (8)-(10), in order to calculate the end-to-end function $F(Y)$, it is required to learn to obtain the parameter $\theta = \{W_1, W_2, W_3, B_1, B_2, B_3, \}$ of the convolutional neural network. If there is a training set consisting of a large number of HR images $\{X_1\}$ and its corresponding LR image $\{Y_1\}$,θ could be obtained by interpolation between the minimized $F(Y_1, \theta)$ and the original HR image X_1 . The SRCNN algorithm uses the mean square error(MSE) as the loss function:

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \|F(Y_i; \theta) - X_i\|^2 \tag{11}$$

Where n is the number of training samples. SRCNN uses the standard gradient descent method to minimize the randomness of the loss function.

3.2. **Proposed methodology.** Time-series remote sensing images are repeated observations of the same region. The overall structure of the image will not change greatly during a period of time, but the images of different phases will have changes in spectral and spatial information. This is mainly caused by two factors: on the one hand, owing to natural objects are affected by phenological effects, such as vegetation features will show different growth states during different crop growth periods, resulting in changes in spectral information; On the other hand, it is a change in the type of ground objects. This is mainly due to the changes in the natural surface caused by artificial construction or sudden natural phenomena. It can also be reflected by changes in spectral information and spatial information on remote sensing images. Different from the reconstruction of image information with the completely different overall structure in the super-resolution of a natural image, the reconstruction of the predicted time-phase remote sensing image is mainly to capture change information of the image and consider the temporal variation of the remote sensing image, in this paper, the difference between remote sensing images in different time periods is used to represent the temporal change. Combined with the characteristics of SRCNN and sparse representation algorithm, the spatio-temporal reconstruction of the fine-resolution reflectance difference image representing the change of spatio-temporal information is realized. Finally, the fine-resolution image of the unknown phase is linearly combined by the fine-resolution image of the known phase and the reconstructed difference image. The main steps of the model can be divided into three steps: remote sensing image SRCNN network construction and training, residual image fine and coarse-resolution joint dictionary training, and fine-resolution reflectivity image reconstruction. The specific process is shown in Fig.3.

(1) Make a difference image between the Landsat images H_1 and H_3 on t_1 and t_2 , and obtain the difference image H_{31} , then resample the simultaneously corresponded MODIS images L_1 and L_3 to the same size as H_1, and then obtain difference image L_{31} .

(2) Using L_{31} and H_{31} as the input to the trained SRCNN network, fine-resolution difference image H^*_{31} can be reconstructed by using the trained SRCNN network. Calculate the difference between H_{31} and H^*_{31} to obtain a residual image H^*_{31} containing details that have not been reconstructed.

(3) Making fine-resolution residual image H^*_{r31} and coarse-resolution difference image L_{31} in patch form (this article uses 8×8), and expand into 64×1 column vector. Forming fine-resolution training sample set $H^*_{r31} = \{h^{*1}_{r31}, h^{*2}_{r31}, ..., h^{*n}_{r31}\}$ and low- resolution training sample set $L_{31} = \{l^1_{31}, l^2_{31}, ..., l^n_{31}\}$.

(4) Randomly extract 1024 column vectors of corresponding positions from the fine- and coarse-resolution sample sets as the initial joint dictionary of fine and coarse resolution. The OMP algorithm [21] is used to sparsely optimize the initial joint dictionary, and the sparse representation coefficient a is obtained. Utilizing the K-SVD algorithm [20] to update the joint dictionary to obtain the fine-resolution dictionary D_h and the coarse-resolution dictionary D_1 .

(5) The low-resolution image on t_1 and t_3 are upsampled to the spatial resolution of the fine-resolution image, and the coarse-resolution difference image L_{21} between t_1 and t_2 is obtained. Input L_{21} into the trained SRCNN network and output the estimated fine-resolution difference image H^*_{21} .

(6) Divide L_{21} and H^*_{21} into 8×8 image patches and expand the image patch into 64×1 column vectors. Obtain fine-resolution difference initial estimate image patch vector h^*_{21} and coarse-resolution training sample vector l_{21} .

(7) Estimate the sparse representation coefficient a of each l_{21} for the low resolution dictionary D_i by the OMP algorithm. Since the sparse representation coefficients of fine- and coarse-resolution dictionary are the same during the joint dictionary update process, the corresponding high resolution image patch h_{21} can be reconstructed by the formula $D_h a$.

(8) The h^*_{21} obtained in step (6) and the h_{21} obtained in step (7) are linearly superimposed, and step (7) is repeated until the image-by-image patch operation is completed. The image patches are superimposed and restored into an image, and the overlap between the image patches is averaged.

Thus obtaining predicted fine-resolution difference image H^p_{21} between t_1 and t_2 . Similarly, the predicted fine-resolution difference image H^p_{32} between t_2 and t_3 can be obtained. Finally, the predicted fine-resolution image H^p_2 can be obtained by formula (3).

4. **Simulation experiments and results analysis.** The geographic coordinates of the center of the experimental area are 45° north latitude and 126° east longitude. The experimental area has a temperate continental climate and it is an important commercial grain base in China with a vast territory and abundant resources. The type of land cover is mainly farmland. These characteristics facilitate the development of our experiment. This paper uses the Landsat8 OLI images of the experimental area and the corresponding MODIS images as experimental data. The Landsat8 OLI images use the ENVI-Flassh

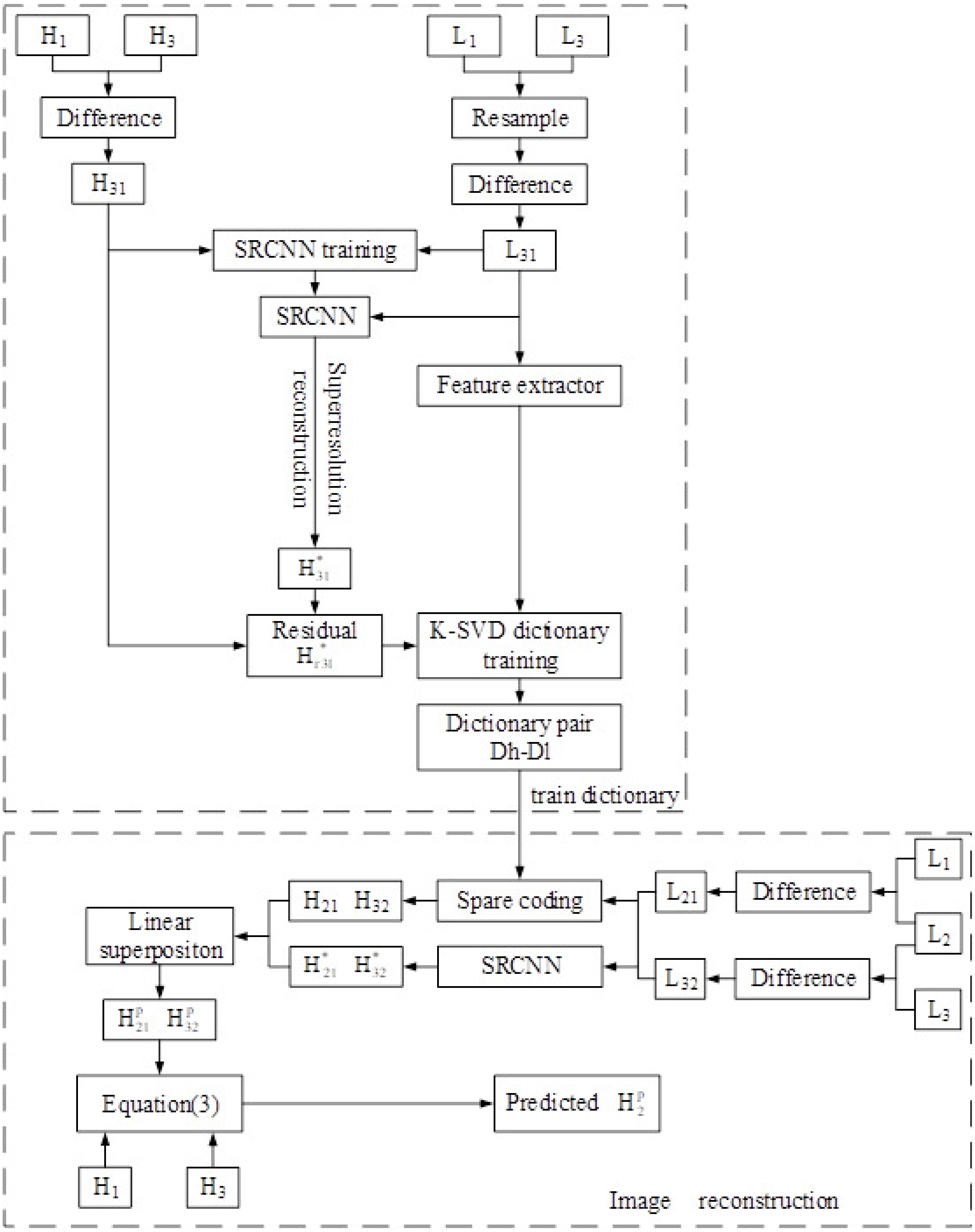
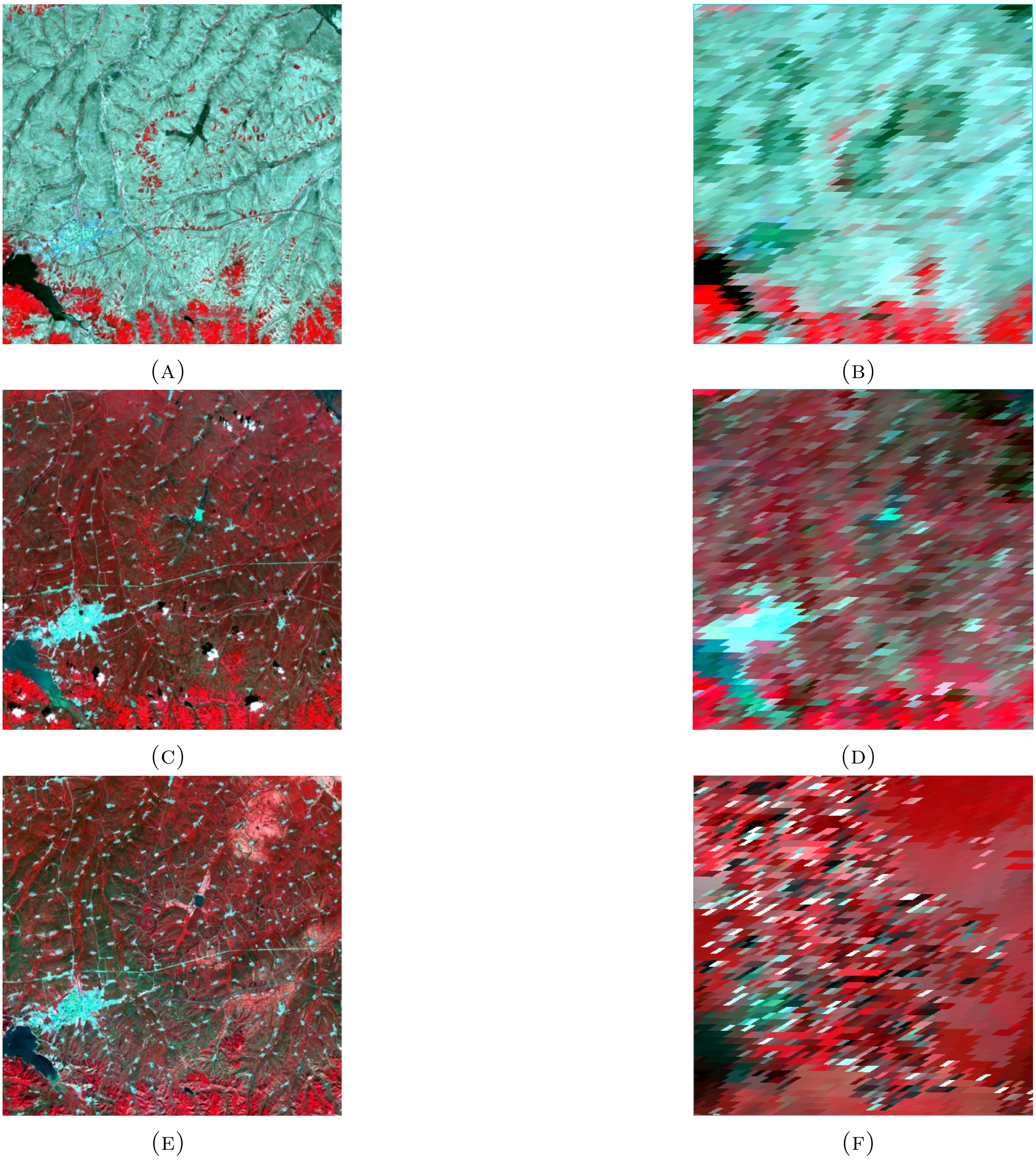


Figure 3. The experimental flow chart is divided into two parts: train dictionary and image reconstruction.

Atmospheric Correction Module to achieve atmospheric correction. Geometrically corrected atmospherically corrected images were geometrically corrected using a 1:10000 topographic map and re-projected to UTM-WGS84. The coordinate system has an error of fewer than 0.5 pixels. The pixel area (with a spatial resolution of 30 m) was intercepted as an experimental analysis area. The MODIS reflection raw data is projected by a sinusoidal projection method. Reproject the MODIS image into the same UTM-WGS84 coordinate system as the Landsat 8 OLI image by using MODIS Reprojection Tools. The surface coverage type of this area is stable, and the change in surface reflectance is considered to be determined only by the phenological phenomena of the vegetation. Fig.4 (A) and (B) show the Landsat and MODIS images on May 29 in 2015 respectively, Fig. 4 (C) and (D) show the Landsat and MODIS images of the experimental area on June 27 in 2015 respectively, and Fig.4 (E) and (F) show the Landsat and MODIS images of the experimental area on August 9 in 2015 respectively. Both are standard false-color images, select bands 5, 4, 3 for Landsat8 OLI images and select bands 2, 1, 4 for MODIS images, the combination of NIR, Red, and Green bands. It can be seen that the MODIS and Landsat images on the same day are very similar, and both images clearly reflect the change in surface reflectance during this period. In this section, we compare the proposed method with the well-known STARFM algorithm and SPSTFM algorithm by using the Landsat 8 OLI images and MODIS images of the experimental area. The predicted image is then compared to the actually observed image, and subjective evaluations and goals are evaluated to assess prediction accuracy. Root mean square error (RMSE), the mean absolute difference (AAD), exponential-structural similarity (SSIM) ‘[22] and ERGAS [23] were chosen as objective quality evaluation indicators.

Parameter selection of SRCNN: the initial values of the three-layer filter $f_1 \times f_1$, $f_2 \times f_2$ and $f_3 \times f_3$ are 9×9 , 1×1 , 5×5 ,n_1 is 64, n_2 is 32, respectively. Due to the influence of the learning rate of the end convolutional layer on the convergence, the learning rate of the first two layers is $\eta_1 = \eta_2 = 10^{-4}$ and the last layer is $\eta_3 = 10^{-5}$. In order to compare and analyze the advantages and disadvantages of different methods, the Landsat and MODIS remote sensing images of May 29, 2015 and August 9, 2015 are used as two sets of base time images, and the Landsat remote sensing image on June 27, 2015 is used as a reference. The images were compared and analyzed in three models: STARFM, SPSTFM, our method.



(A)

(B)

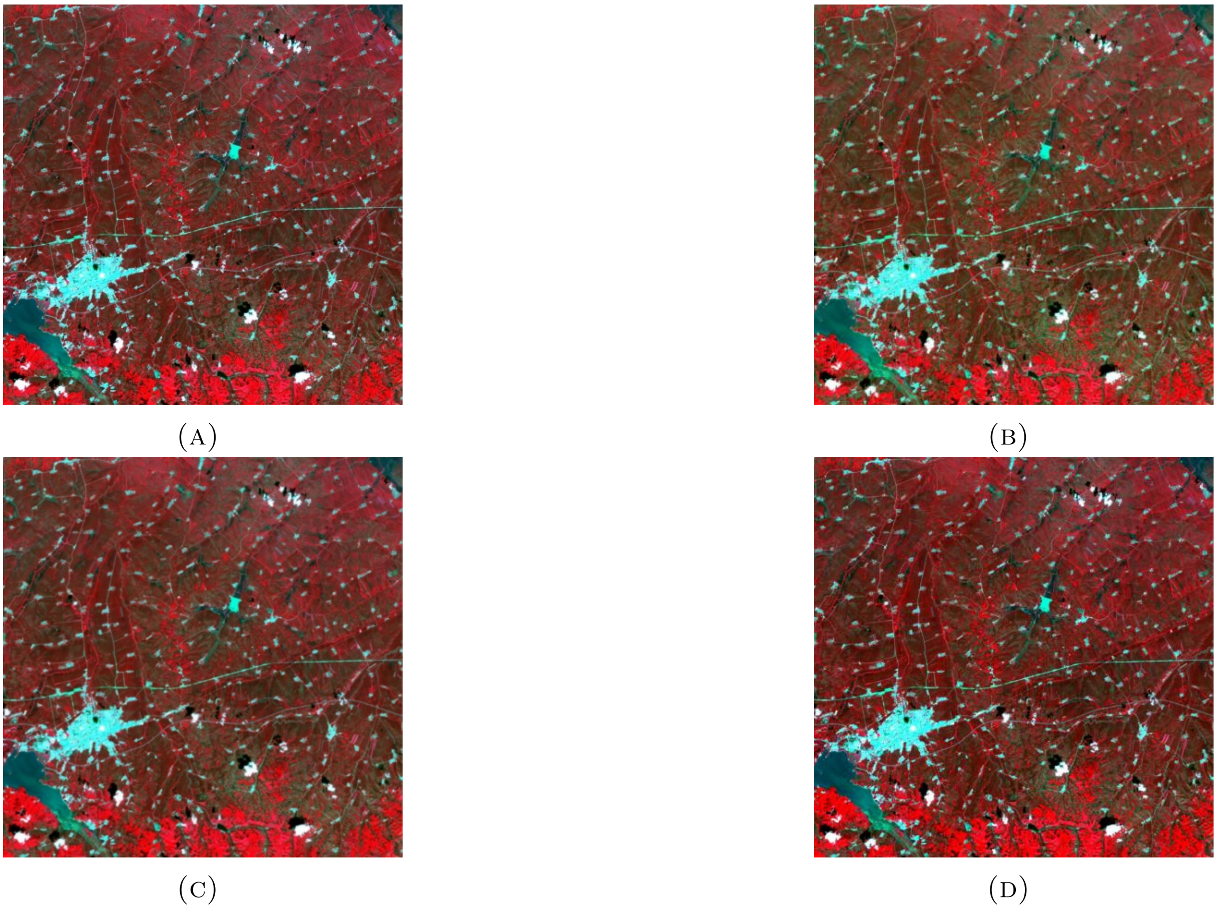
(C)

(D)

(E)

(F)

FIGURE 4. Landsat image and MODIS image of the study area. (A) Landsat image on May 29, 2015. (B) MODIS image on May 29, 2015. (C) Landsat image on June 27, 2015. (D) MODIS image on June 27, 2015. (E) Landsat image on August 9, 2015. (F) MODIS image on August 9, 2015.



(A)

(B)

(C)

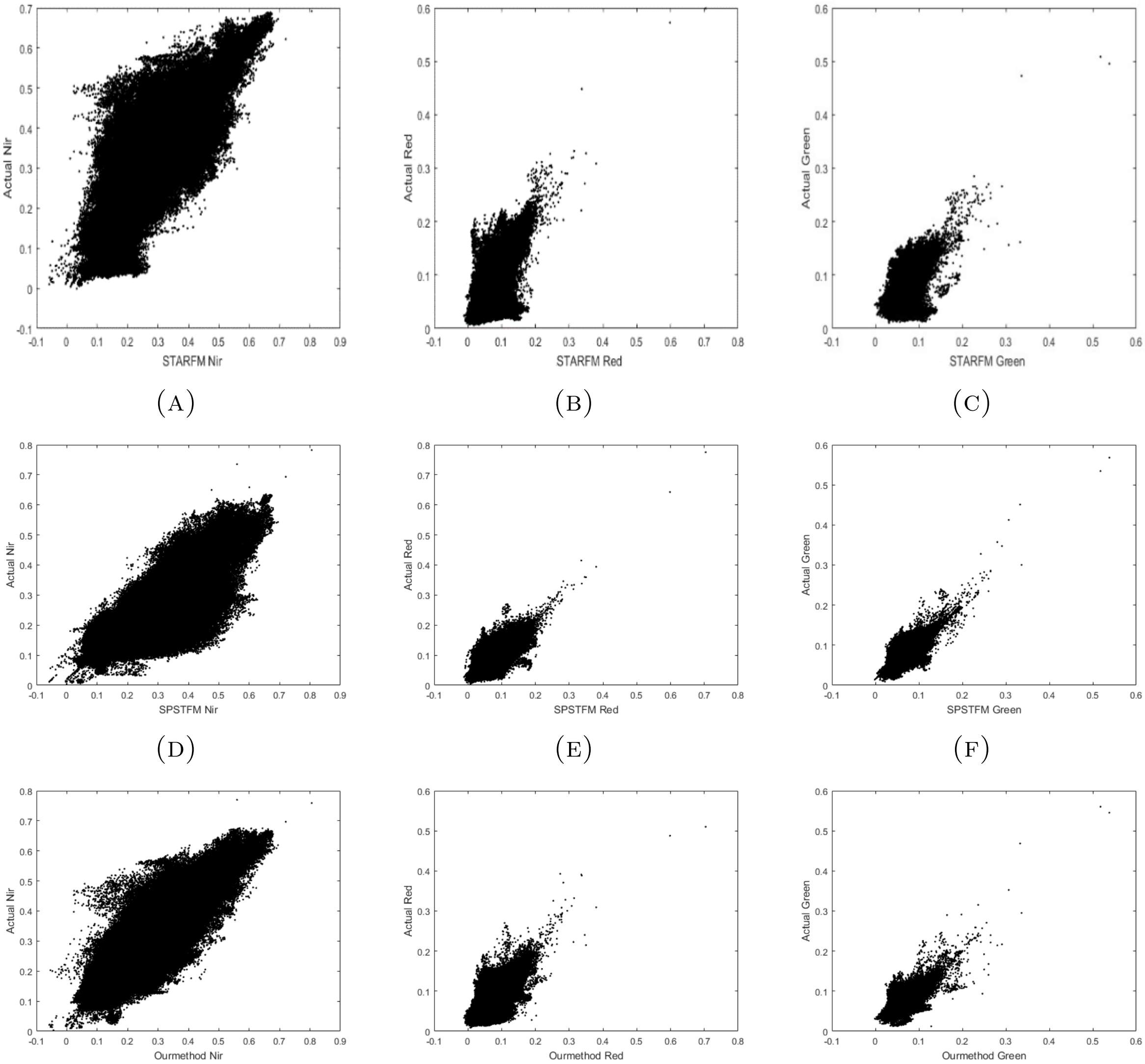
(D)

Figure 5. Comparisons between actual and predicted surface reflectances with focus on seasonal changes. (A) is actual surface reflectance. (B) is predicted surface reflectance using STARFM. (C) is predicted surface reflectance using SPSTFM. (D) is predicted surface reflectance using our method5.

Table 1. Accuracy evaluation of fusion results in experimental areas

Fusion method	AAD			RMSE			SSIM			ERGAS
	Red	Green	Nir	Red	Green	Nir	Red	Green	Nir	
STARFM	0.0179	0.0145	0.0117	0.0238	0.0207	0.0126	0.791	0.770	0.823	1.4578
SPSTFM	0.0160	0.0139	0.0092	0.0199	0.0171	0.0092	0.820	0.782	0.832	1.0926
Our method	0.0153	0.0132	0.0090	0.0175	0.0156	0.0077	0.842	0.791	0.841	0.9527

From Fig.5, we can see that compared with the proposed method and SPSTFM algorithm, the fusion result of the STARFM algorithm has a large fusion error, mainly because the method searches for similar pixels near the center pixel based on the reference time. It is assumed that adjacent similar pixels in the two basic phase periods experience a similar phenological change with the center pixel, and the reflectance of the similar pixels is used to obtain the reflectance of the center pixel of the predicted phase. When large phenological changes occur, due to the high heterogeneity of the selected regions, the central pixel and adjacent similar pixels have different degrees of phenological changes in the cycle, which results in inaccurate fusion results. The comparisons in terms of AAD, RMSE, ERGAS, and SSIM are listed in Tab.1, the average AAD values of the three bands



(A) (B) (C)

(D) (E) (F)

(G) (H) (I)

Figure 6. Scatter plots of the predicted reflectance against the actual reflectance for NIR-red-green bands from left to right. (A)-(C) are scatter plots of the predicted reflectance by using STARFM against the actual reflectance in the NIR, red, and green bands, respectively. (D)-(F) are scatter plots of the predicted reflectance by using SPSTFM against the actual reflectance in the NIR, red, and green bands, respectively. (G)-(I) are scatter plots of the predicted reflectance by using our method against the actual reflectance in the NIR, red, and green bands, respectively..

for STARFM, SPSTFM, Our method are 0.0147, 0.0130, 0.0125, and 0.0125,respectively, and the average RMSE values of the three bands for these models are 0.0190, 0.0154, 0.0136,respectively.These indicate that our method can reconstruct the Landsat surface reflectance more precisely than STARFM and SPSTFM. The average SSIM values of the three bands for these methods are 0.795, 0.811, 0.825, respectively, and this indicates that our method can retrieve more precise structural details with smaller reflectance deviations on the surface reflectance than STARFM and SPSTFM. The ERGAS values for these methods are 1.4576, 1.0926, and 0.9527, respectively, and this indicates that the spatial details and spectral colors of our fusion result are better than this of STARFM and SPSTFM.

5. Conclusions.

This paper proposes a spatio-temporal fusion model based on sparse representation and SRCNN. Using our method to compare with STARFM and SPSTFM, the results show that our method has better fusion precision and better spatial detail reconstruction.

Spatio-temporal fusion is the complementary advantage of remote sensing image information. In the future research, on the one hand, how to make full use of remote sensing image information to make it more widely used in agricultural monitoring and other fields; on the other hand, in order to explore the spatial information of low -resolution image MODIS more widely, more and more reasonable training sample sets are extracted to train the SRCNN network to obtain better Landsat prediction results.

Acknowledgment. This work is supported by National Natural Science Foundations of China (61871150), National Key R&D Program of China (2016YFB0502502).

REFERENCES

[1] F. Gao, J. Masek, M. Schwaller, et al. On the blending of the Landsat and MODIS surface reflectance: predicting daily Landsat surface reflectance *IEEE Transactions on Geoscience & Remote Sensing*,44(8): 2207-2218, 2006

[2] X. L. Zhu, J. Chen, F. Gao, et al. An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions, *Remote Sensing of Environment*, 114(11): 2610-2623, 2010.

[3] T. Hilker, M. A. Wulder, N. C. Coops, et al. A new data fusion model for high spatial- and temporal-resolution mapping of forest disturbance based on landsat and MODIS *Remote Sensing of Environment*, 113(8): 1613-1627, 2009.

[4] M. Q. Wu, Z. Niu, C. Y. Wang et al. Use of MODIS and Landsat time series data to generate high-resolution temporal synthetic Landsat data using a spatial and temporal reflectance fusion model, *Journal of Applied Remote Sensing*, 6(1): 063507-1-063507-13, 2012.

[5] F. Q. Zhang, T. Y. Wu, J. S. Pan, G. Y. Ding, Z. Y. Li. Human Motion Recognition Based on SVM in VR Art Media Interaction Environment, *Human –centric Computing and Information Sciences*, 9(1):1-15, 2019.

[6] F. Q. Zhang, T. Y. Wu, G. Y. Zheng. Video salient region detection model based on wavelet transform and feature comparison, *EURASIP Journal on Image and Video Processing*,2019(1):1-10, 2019.

[7] B. Huang, H. Song. Spatio-temporal Reflectance Fusion via Sparse Representation, *IEEE Transactions on Geoscience & Remote Sensing*, 50(10): 3707-3716, 2012.

[8] H. Song, B. Huang. Spatio-temporal satellite image fusion through one-pair image learning. Geoscience and Remote Sensing, *IEEE Transactions on*, 51(4): 1883-1896, 2013.

[9] Eric Ke Wang, Chien-Ming Chen, Mohammad Mehedi Hassan, et al. A deep learning based medical image segmentation technique in Internet-of-Medical-Things domain *Future Generation Computer Systems*, 108:135-144, 2020.

[10] Eric Ke Wang, Chien-Ming Chen, Fan Wang, et al. Joint-learning segmentation in Internet of drones (IoD)-based monitor systems, *Computer communications*,152 :54-62, 2020.

[11] D. Yu, L. Deng. Deep learning and its applications to signal and information processing, *IEEE Signal Processing Magazine*, 28(1):145-154, 2011.

[12] F. H. C. Tivive, A. Bouzerdoum. A New Class of Convolu-tional Neural Networks (SICoNNets) and Their Appl-ication of Face Detection, *Proceedings of International Joint Conference on Neural Networks.Washington D.C.,USA: IEEE Press*, : 2157-2162, 2003.

[13] A. Krizhevsky, I. Sutskever, G. E. Hinton. Image Net classification with deep convolutional neural networks, *Proceedings of the 2012 Advances in Neural Information Processing Systems. Lake Tahoe, Nevada, USA: Curran Associates, Inc*, 1097−1105, 2012.

[14] E. K. Wang, X. Zhang, F. Wang, T. Y. Wu, C. M. Chen. Multilayer Dense Attention Model for Image Caption, *IEEE Access*, 7:66358-66268, 2019.

[15] Girshick, Ross, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, *CVPR IEEE*, 2014.

[16] C. Dong, C. C. Loy, K. He, et al. Image super-resolution using deep convolutional networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2): 295-307, 2016.

[17] E. K. Wang, F Wang, S Kumari, J. H. Yeh, CM Chen. Intelligent monitor for typhoon in IoT system of smart city, *The Journal of Supercomputing*,:1-20, 2020.

[18] K. K. Tseng, R. Zhang, C. M. Chen, M. M. Hassan. DNetUnet: a semi-supervised CNN of medical image segmentation for super- computing AI service, *The Journal of Supercomputing*,:1-22, 2020.

[19] J. Yang, J. Wright, T. Huang, and Y. Ma. Image super-resolution via sparse representation, *Image Process*, 19(11): 2861–2873, 2010.

[20] M. Aharon, M. Elad, and A. M. Bruckstein.The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation, *Signal Process*,54(11):4311-4322, 2006.

[21] G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *Constructive Approx*,13(1):57-98, 1997.

[22] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity, *Image Process*,13(4):600-612, 2004.

[23] M. M. Khan, L. Alparone, and J. Chanussot. Pansharpening quality assessment using the modulation transfer functions of instruments, *Geosci. Remote Sens*,47(11):3880–3891, 2009.