

Accurate Patch Based Satellite Image Fusion Using Deep Convolution Neural Network

P. Veena, C. C. Manju

Lecturer in Computer Engineering, NSS Polytechnic College, Pandalam, Pathanamthitta, Kerala-689501

Lecturer in Computer Engineering, Government Polytechnic College, Punalur

Abstract: Spatiotemporal fusion can combine Landsat and MODIS photos, which have complimentary spatial and temporal features, to create high-resolution data. A deep convolutional neural network (DCNN)-based spatial fusing method is presented in this work to manage vast remote sensing data in useful applications. Landsat image areas are derived from low-resolution MODIS photos. A large number of training patches are previously collected with respect to different attributes such as color, edge and pixel statistics. Further the images are patched by 3x3 block size and converted to array to perform the DCNN training process. Regression based DCNN training algorithm is performed to predict the missing information from the local patches. The trained DCNN module is further used to generate the predicted output using the same procedure used for the training process. In the testing, finally the patches are converted into matrix to obtain the complete image output. Two standard files from Landsat-MODIS are extensively evaluated. Images are assessed using RMSE, MAE, RMAE, and MACE metrics. The proposed strategy yields more precise fusing results than sparse representation-based methods. From the execution of this work an average of RMSE, MAE, RMAE and MACE of 0.004, 0.09, 0.02, and 4.65 respectively achieved for multiple images from dataset.

Keywords: Spatiotemporal, remote sensing, MODIS images, deep convolutional neural network.

I. INTRODUCTION

Huge quantities of remote sensing pictures can now be acquired on a daily basis thanks to the quick development of sensor networks for remote sensing. This significantly restricts the potential applications of remote sensing pictures and is a result of the technological constraints of the sensors as well as other variables. Fortunately, remote sensing image fusing is an efficient method that can be utilized to

acquire superior high spatio-temporal- and spectral-resolution pictures by combining the information that is complimentary to one another. It is described as the synergistic combination of two or more picture data sets, the goal of which is to produce a knowledge of the phenomena under investigation that is superior to the knowledge that can be achieved from the knowledge obtained from individual data sets.

The subject of research known as image fusion has been around for quite some time and has seen significant advancements in recent years. The growing demand for satellite pictures with better geographic, temporal, and/or spectrum resolution is the driving force behind these advancements. In the past, techniques of image fusing were primarily focused on improving spatial resolution and integrating pictures from multiple input modalities. Recent developments in this field have shifted the emphasis of these techniques toward the fusion of pictures with high temporal frequency and fine spatial resolution. The previous few decades have seen the development of a wide variety of instruments that are carried by satellites. For example, the Moderate resolution Imaging Spectroradiometer (MODIS) instrument onboard the Terra and Aqua spacecraft of NASA captures data spanning an area of 2330 kilometers at various spatial (250 m, 500 m, and 1 km) and chronological resolutions. These satellites have been in operation since 1999 and 2002, respectively.

This can be accomplished by combining the two sets of data. In general, one can divide spatiotemporal fusing into the following four models: 1) models based on transformations; 2) models based on image reconstructions; 3) models based on Bayesian statistics; and 4) models based on learning. In order to

conduct fusion and acquire high-resolution data at uncertain periods, the original picture pixels are transferred to an artificial space using the data transformation model.

Both of these models and algorithms are adapted to account for changing reflectance over time. (STAARCH). Models that are founded on Bayesian theory apply the Bayesian statistical principle to quantitative statistics. Bayesian-based models offer a number of benefits when it comes to the processing of a variety of incoming pictures to generate improved projection results. On the other hand, the majority of the conventional algorithms described above depend on conditions that are predetermined in advance.

When using a learning-based fusion model, there is typically no requirement to explicitly create fusion principles. It has the capability to automatically learn the finest fundamental characteristics from a variety of high-quality input datasets and produce combined pictures of a high-quality. Sparse representation and deep learning technology are the two primary approaches that are utilized in the construction of models that are based on learning at this time.. Through the association between these images, the researchers are able to extract some important feature information. The high-resolution pictures that are needed for projection are reconstructed by the algorithm. Although these methods are capable of producing superior fusing outcomes than more conventional methods, their ubiquitous applicability is restricted due to a number of constraints, such as patchy coding, high computational cost, and computational complexity.

Deep learning is a technique that replicates, to a large extent, the functioning characteristics of the neural structure found in the human brain. These characteristics include the continuous transmission of information between the various neurons. There is no one right method to construct an infrastructure for a deep learning network. Convolutional neural networks (CNNs) are becoming increasingly popular as a lightweight and effective method for the extraction of picture features and the reconstitution of images that also possesses a powerful capacity for learning.

The following is an outline of this paper's primary segments in chronological order: In Section II, the

pertinent resources and the suggested technique are both described. The third section details a number of experiments, along with the outcomes of those experiments and an analysis of the findings. In Section IV, we examine the effectiveness of the suggested network structure on various datasets, as well as the benefits of using such a structure. In Section V, a summary of the entire paper's findings is presented, and a prognosis for future research is discussed.

II. LITERATURE SURVEY

Huimin and coworkers proposed a multiscale-attention STF network design. (MANet). The expanded convolution-based multiscale mechanism module is responsible for the extraction of particular characteristics from remote sensing pictures with a poor spatial resolution at multiple scales. The channel thought process easily grows accustomed the substantial opposes of the power source feature map channels to retain greater archaeological and geographical understanding in the further collection manipulative, the non-local attention process adjusts the initially fusion images to obtain more precise suggested images by establishing the correlation between pixels. In the experiments, we make use of two separate and distinct databases. The machine learning-based and deep learning-based fusing techniques are outperformed by the MANet approach, which requires fewer inputs.

Weisheng et al. proposed a convolutional neural network-based spatial merging method. This network received focus and multiscale processes. (MANet). Propose spatial merger with one set of images. These three steps cover most of the work. First, get feature maps from two different-scaled images. Use the focusing device to focus on the most important feature map information. Third, the image must be rebuilt. We used two standard datasets and three state-of-the-art spatial fusion methods to compare our results. Our method yields more regional detail and more precise time projections than prior methods.

A Convolutional Neural Network was used as the foundation for the Multi-scene Spatiotemporal Fusion Network (MUSTFN) technique that was developed by Peng et al. (CNN). Our method combines pictures taken by numerous cameras, each of which have a

unique resolution. To do this, we make use of multi-level characteristics. In addition to this, MUSTFN makes use of the multi-scale characteristics in order to mitigate the impacts of geometric registration mistakes that occur between various pictures. The Landsat-7 pictures with a resolution of 30 meters and the MODIS images with a resolution of 500 meters can be fused using our suggested technique, MUSTFN, which performs significantly better than several of the methods that are currently in use. This was accomplished by combining the data from the two sets of images. When combining long-term Landsat-8 composite pictures with MODIS images over a broad territory, we finally demonstrated the applicability of MUSTFN by achieving an average rMAE of 9.18 percent. In general, the findings of this research point to the viability of MUSTFN as a solution to the difficulties associated with image fusing.

Deep learning was the foundation for an innovative approach to spatiotemporal integration that was suggested by Huang and colleagues. A focus multistate feature fusion network are used to create a network that statistically analyzes inner picture features for various input image characteristics. Combining a focus mechanism and a multistate feature fusion network created this network. Spatiotemporal fusion includes a multiscale feature fusing tool. We add a new edge loss function to the compound loss function. This lets us blend images with more edge information. In comparison to the current mainstream spatiotemporal fusion techniques, our suggested model achieves outstanding results on both datasets in terms of index performance as well as picture features. These results can be found by clicking here.

Classical spatio-temporal fusion algorithms such as STARFM and SPSTFM, as suggested by Yang and Wang, will have significant fusion mistakes when phenological changes or type changes occur, according to their research. We suggested a novel

method for the integration of spatiotemporal information based on the spatial feature information of the picture. This method integrates SRCNN, which stands for "Super-Resolution Convolutional Neural Network," with sparse representation. After combining SRCNN and sparse representation in order to finish the feature reconstruction of the reflectance change image, the reconstructed image is then superimposed by the time weight in order to acquire the anticipated reflectance image. Experiments demonstrate that the suggested technique is superior to the well-known spatio-temporal fusing algorithms known as STARFM and SPSTFM.

The main objectives of this work is

- To improve novel spatiotemporal fusion approach with DCNN.
- To improve the performance of the image.
- To reduce the computational complexity and reduce the error of the image.

III. PROPOSED METHOD

Image fusion can improve the quantity of the information present in the images. In this work Landsat and MODIS images are fused to generate the high-resolution image output. Generally, the MODIS images has low resolution content when compared to the Landsat images. Due to high data rate, there is probability to loss the information while acquiring the image from sensor. Fig.1 shows the proposed block diagram for image fusion using deep convolution network with patching techniques. Initially the MODIS images are up-sampled to the size of the landsat images. Further it will be patched and converted to array to perform the training and testing process. DCNN based regression algorithm is used to predict the new information based on the previous samples. The images are formed after prediction by combining small patched to images.

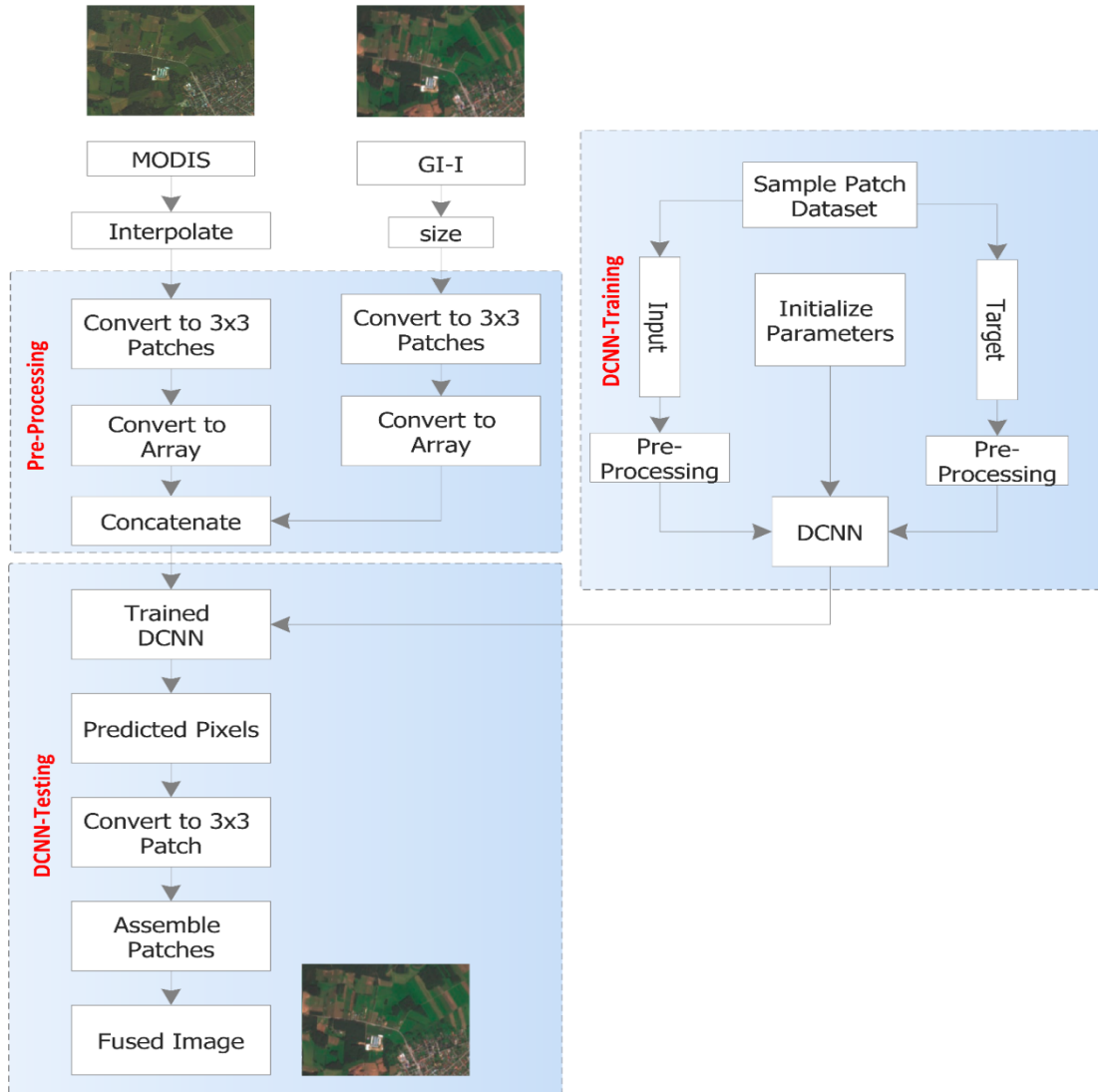


Fig.1. Block Diagram of proposed method

A. Image interpolation

At some point during the process of creating a digital photograph, image transformation will take place. This may take place during the bayer demosaicing step or during the enhancement step. This occurs whenever you transform your picture from one pixel grid to another, whether you are resizing it or mapping it. The linear interpolation method is further developed into

$$P(x, y) = P(x_1, y_1) * W(x_1, y_1) + P(x_1, y_2) * W(x_1, y_2) + P(x_2, y_1) * W(x_2, y_1) + P(x_2, y_2) * W(x_2, y_2)$$

(1)

Where, W, represents the weight (or area) assigned to the pixel samples belonging to the location (x_1, y_1) , (x_1, y_2) , (x_2, y_1) , (x_2, y_2) .

B. Deep Convolutional Neural Network (DCNN)

The input layer, the convolution layer, the pooling layer, the full-connection layer, and the output layer

the bilinear interpolation method. The foundation of this method is the execution of interpolation in both directions to make an educated guess or approximation regarding the missing pixel value at $P(x,y)$, The value of the pixel that is located at the coordinates (x, y) in the destination picture is represented by the "middle" sample, P.

are the primary components that make up the fundamental framework of the deep CNN.

Input picture is received and stored in matrix form by the input layer, which is the function of the input layer. Assuming that the construction of the retrograde deep CNN is a L layer, then x^l represents the feature of No. l layer, $l = 1, 2, 3, \dots, L$, $x^1 = \{x_1^1, x_2^1, x_3^1\}$ where, x_1^1, x_2^1 , and x_3^1 , represents the information contained in the red, green, and blue channels, in that order. The extraction of characteristics via the use of the convolution procedure is the responsibility of the convolution layer. The capability of the regressive deep CNN to articulate features will be improved if the design is done correctly, and this improvement will occur in tandem with an increase in the number of convolution layers. It is possible to compute the feature graph of the first convolution layer using the following formula:

$$x_j^l = f(\sum_{i=1}^{j-1} G_{i,j}^l (k_{i,j}^l \otimes x_i^{l-1}) + b_j^l) \quad (2)$$

Where $k_{i,j}^l$ and b_j^l are the weights of the convolution kernel and biases, $G_{i,j}^l$ is the connection matrix convolution layer and the feature graph of the previous convolution layer; the symbol \otimes defines the convolution operation.

The pooling layer's job is to cut down on the feature dimensions, so that's what it does. In most cases, the pooling layer comes after the convolutional layer, and the pooling procedure has the ability to preserve some degree of spatial invariance. The feature graph x_j^l of the pooling operation in the l layer can be calculated as

$$x_j^l = p(x_j^{l-1}) \quad (3)$$

Where $p(x)$ represents the pooling operation.

The completely connected layer's job is to take the deep feature that was extracted from the layers that came before it and turn it into a feature vector. As a result, this layer is typically placed behind the layer that is responsible for feature separation. It is possible to derive the formula for computing the feature vector XL in the completely connected layer as

$$x^l = f(w^l x^{l-1} + b^l) \quad (4)$$

Where w^l represents the joining weight amongst two adjacent network layers and b^l is the offset and $f(x)$ is the activation function. Table.1. shows the parameter specification of DCNN. The DCNN needs to Loss function of Mean squared error, Epochs of 50, weight decay of 0.0001, Momentum of 0.9, Initial learning rate of 0.01, Interpolation method of Bicubic, Size of training batches of 64.

Table 1. Parameter specification of DCNN

Parameter	Value
Loss function	Mean squared error
Epochs	50
weight decay	0.0001
Momentum	0.9
Initial learning rate	0.01
Interpolation method	Bicubic
Size of training batches	64

IV. RESULTS AND DISCUSSIONS

A. Dataset

The Coleambally Irrigation Area (CIA) in southern New South Wales, Australia, was the first study site. From 2001 to 2002, CIA had 17 cloud-free Landsat–MODIS image-pairs for austral summer growth season. CIA dataset time dynamics are crop phenology over a single watering season. Small area sizes make CIA spatially diverse. LGC is more temporally active due to a mid-December 2004 storm that inundated 44% of the site. The CIA dataset used Landsat-7 ETM+ photos atmospherically adjusted by MODTRAN4 [30]. Landsat-5 TM photos were atmospherically adjusted for the LGC file. Pre-processing Landsat data used the Australian Geodetic Datum (AGD66) for geo correction. CIA and LGC datasets have 25 m spatial precision and 2040 × 1720 and 2720 × 3200 image sizes, respectively. Both study sites use Terra MOD09GA Collection 5 MODIS photos with 500 m spatial resolution. The closest neighbor method up-sampled MODIS photos to 25 m to match Landsat data. Maximizing the correlation function between Landsat and MODIS images yielded an optimum offset for each MODIS image to co-register them with sub-pixel accuracy. We used Landsat and MODIS bands 1, 2, 3, 4, 5, and 7 for experiments. We modified MODIS photos to fit Landsat images due to their distinct band order layouts.

B. Evaluation metrics

The best metric for comparing spatial fusion models is spectrum accuracy. RMSE, MAE, and R2 are common spectrum indices. The relative Mean Absolute Errors (rMAE) can also assess the ratio of mean absolute errors to true values. rMAE is one of the most reliable predictors of spectrum variations between expected

and real values. Results with rMAE below 15% are better. MAEC was used to assess MAE shift in various conditions. Spatial accuracy helps evaluate fusing results along with spectrum accuracy. The Edge meter is a solid spatial accuracy metric, so compare the mean

difference between the reference image and fused image to assess spatial accuracy. Two small-area tests used the Edge meter because we focused on spectrum fusion results. Calculations:

$$RMSE = \sqrt{\sum_{i=1}^n (F_i - R_i)^2 \frac{1}{n}} \tag{5}$$

Where n is the total number of pixels in the image, R_i is the reference image and F_i is the fused image.

$$MAE = \sum_{i=1}^n |F_i - R_i| \frac{1}{n} \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (F_i - R_i)^2}{\sum_{i=1}^n (F_i - \mu_R)^2} \tag{7}$$

Where μ_R represents the mean value of the reference image

$$rMAE = \frac{1}{n} \sum_{i=1}^n |F_i - R_i| \frac{1}{R_i} \tag{8}$$

$$MAEC = \frac{|MAE_2 - MAE_1|}{MAE_1} \tag{9}$$

$$Edge_{F_{i,j}} = |F_{i,j} - F_{i+1,j+1}| + |F_{i,j+1} + F_{i+1,j}| \tag{10}$$

Where $F_{i,j}$ represents the value of pixels at ith row and jth column in fused image.

C. Performance Analysis

Table.1. shows the performance of proposed method. For the image 1, red channel gets RMSE of 0.08, MAE of 0.63, R^2 of 0.86, rMAE of 17.39,MAEC of 26%, green channel gets RMSE of 0.9, MAE of 0.68, R^2 of 0.79, rMAE of 16.36,MAEC of 33%, blue channel gets RMSE of 0.12, MAE of 0.93, R^2 of 0.82, rMAE of 18.64,MAEC of 16%, For the image 2, red channel gets RMSE of 0.085, MAE of 0.46, R^2 of 0.86, rMAE of 16.34,MAEC of 28%, green channel gets RMSE of 0.089, MAE of 0.48, R^2 of 0.89, rMAE of

17.82,MAEC of 29%, blue channel gets RMSE of 0.02, MAE of 0.23, R^2 of 0.75, rMAE of 17.64,MAEC of 34%. Table.2. shows the comparative Performance of proposed method. This work achieves RMSE of 0.004, MAE of 0.09, R^2 of 0.02, rMAE of 4.65,MAEC of 42% and Edge is -0.05×10^{-2} .

Fig.2 shows the training performance of proposed method Fig.3. shows the (a) input image 1 (b) corresponding MODIS image, (c) Fused Image of input image 1 (d) input image 2 (e) corresponding MODIS image, (c) Fused Image of input image 2.

Table.2. Performance of proposed method

	Channel	RMSE	MAE	R^2	rMAE	MAEC(%)	Edge ($\times 10^{-2}$)
Image 1	Red	0.08	0.63	0.86	17.39	26	-0.06
	Green	0.9	0.68	0.79	16.36	33	-0.03
	Blue	0.12	0.93	0.82	18.64	16	-0.36
Image 2	Red	0.085	0.46	0.86	16.34	28	-0.28
	Green	0.089	0.48	0.89	17.82	29	-0.06
	Blue	0.02	0.23	0.75	17.64	34	-0.09
Image 3	Red	0.36	0.41	0.82	16.52	36	-0.08
	Green	1.36	0.18	0.84	19.46	33	-0.62
	Blue	0.39	0.85	0.78	17.52	27	-0.96

Table.3. Comparative Performance of proposed method

	RMSE	MAE	R^2	rMAE	MAEC	Edge
[8]	1.45	0.99		5.11		-0.06

[10]	0.017	-	-	-	-	-
[11]	1.90	-	-	-	-	-
[12]	0.0059	-	-	-	-	-
This work	0.004	0.09	0.02	4.65	42	-0.05

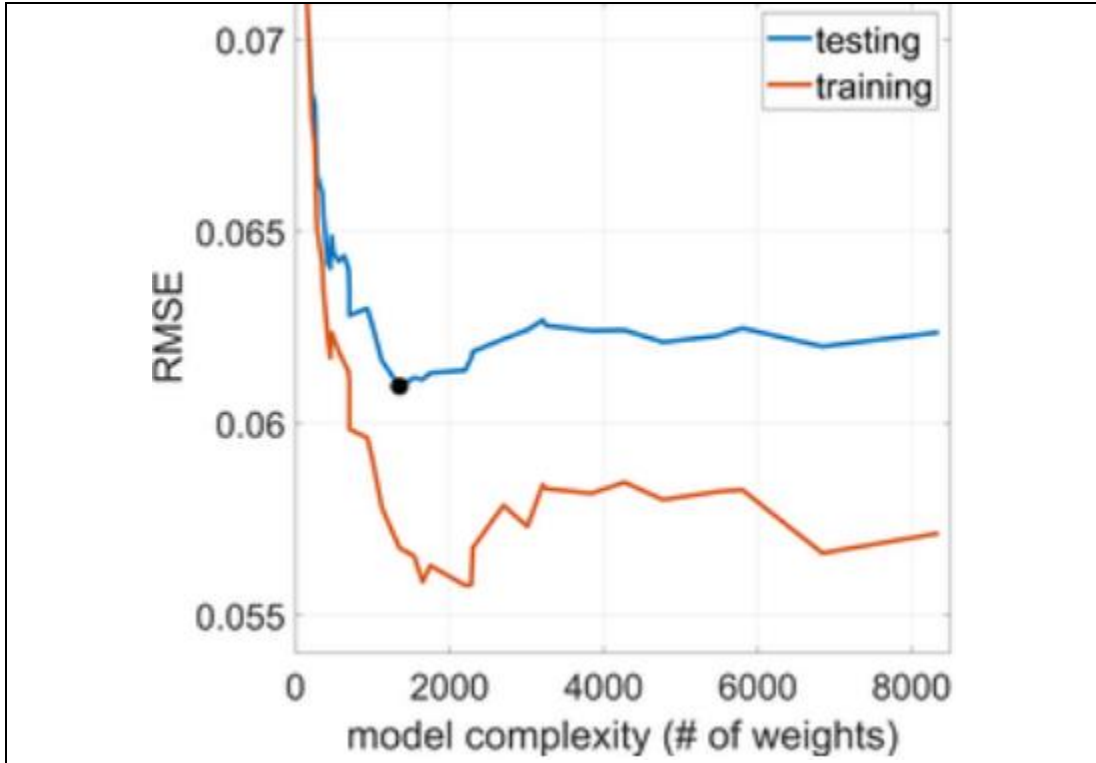
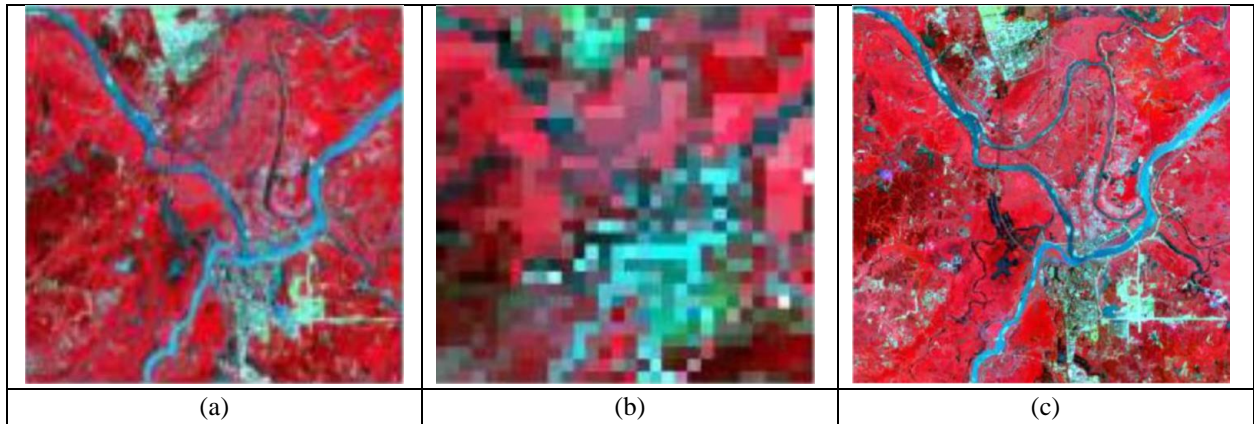


Fig.2 Training performance of proposed method



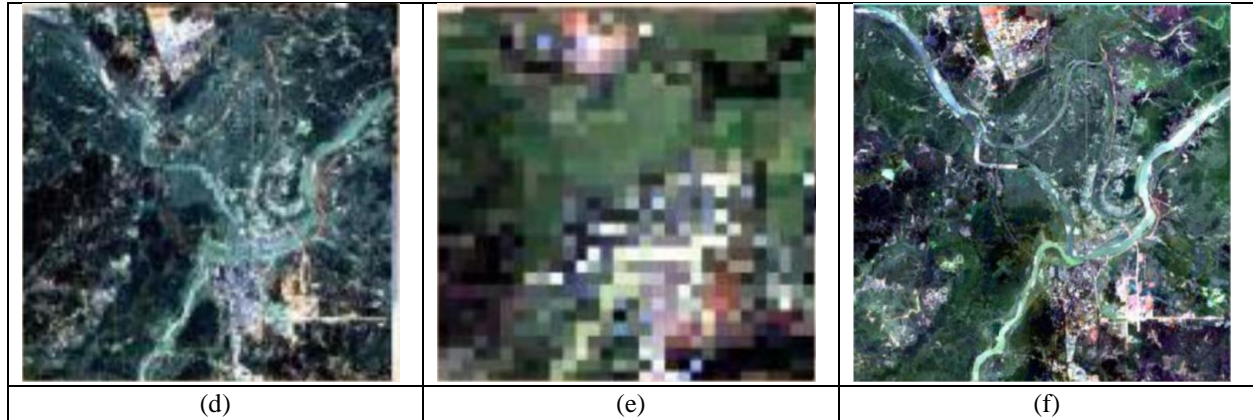


Fig.3. (a) input image 1 (b) corresponding MODIS image, (c) Fused Image of input image 1
(d) input image 2 (e) corresponding MODIS image, (c) Fused Image of input image 2

V. CONCLUSION

In this study, we blended the geographical information from Landsat data with the temporal information from MODIS data to suggest a new technique for spatiotemporal fusion that is based on deep convolutional neural networks (DCNNs). This allowed us to deal with the highly non-linear correspondence relations that exist between MODIS and Landsat data. Because of this, we were able to successfully handle the extremely non-linear correspondence connections. The prediction stage was made up of three levels, each of which contained a DCNN-based prediction phase in addition to a fusion model. These levels were all connected to one another via a fusion model. These learned DCNN models and the fusion model served as the basis for this stage's predictions. The projected images that were produced by the DCNN model were used as intermediate images, and then an HPM module along with a suggestive weighting strategy were used to incorporate the information that was found in prior image pairs. This was done so that the prior information could be completely explored. The suggested method was shown to be superior to other learning-based methods through a series of experimental assessments that were performed on two benchmark datasets. As a result of carrying out this task, an average RMSE, MAE, RMAE, and MACE score of 0.004, 0.09, 0.02, and 4.65 was attained for a number of different pictures taken from the collection.

REFERENCES

- [1]. Shen, Huanfeng, Xiangchao Meng, and Liangpei Zhang. "An integrated framework for the spatio-temporal-spectral fusion of remote sensing images." *IEEE Transactions on Geoscience and Remote Sensing* 54, no. 12 (2016): 7135-7148.
- [2]. Belgiu, Mariana, and Alfred Stein. "Spatiotemporal image fusion in remote sensing." *Remote sensing* 11, no. 7 (2019): 818.
- [3]. Wang, Qunming, Yijie Tang, Xiaohua Tong, and Peter M. Atkinson. "Virtual image pair-based spatio-temporal fusion." *Remote Sensing of Environment* 249 (2020): 112009.
- [4]. Zhu, Xiaolin, Wenfeng Zhan, Junxiong Zhou, Xuehong Chen, Zicong Liang, Shuai Xu, and Jin Chen. "A novel framework to assess all-round performances of spatiotemporal fusion models." *Remote Sensing of Environment* 274 (2022): 113002.
- [5]. Li, Weisheng, Dongwen Cao, Yidong Peng, and Chao Yang. "MSNet: A multi-stream fusion network for remote sensing spatiotemporal fusion based on transformer and convolution." *Remote Sensing* 13, no. 18 (2021): 3724.
- [6]. Cao, Huimin, Xiaobo Luo, Yidong Peng, and Tianshou Xie. "MANet: A Network Architecture for Remote Sensing Spatiotemporal Fusion Based on Multiscale and Attention Mechanisms." *Remote Sensing* 14, no. 18 (2022): 4600.
- [7]. Li, Weisheng, Xiayan Zhang, Yidong Peng, and Meilin Dong. "Spatiotemporal fusion of

- remote sensing images using a convolutional neural network with attention and multiscale mechanisms." *International Journal of Remote Sensing* 42, no. 6 (2021): 1973-1993.
- [8]. Qin, Peng, Huabing Huang, Hailong Tang, Jie Wang, and Chong Liu. "MUSTFN: A spatiotemporal fusion method for multi-scale and multi-sensor remote sensing images based on a convolutional neural network." *International Journal of Applied Earth Observation and Geoinformation* 115 (2022): 103113.
- [9]. Huang, Zhiqiang, Yujia Li, Menghao Bai, Qing Wei, Qian Gu, Zhijun Mou, Liping Zhang, and Dajiang Lei. "A Multiscale Spatiotemporal Fusion Network Based on an Attention Mechanism." *Remote Sensing* 15, no. 1 (2023): 182.
- [10]. Yang, Shuai, and Xiaofei Wang. "Sparse representation and SRCNN based spatio-temporal information fusion method of multi-sensor remote sensing data." *Journal of Network Intelligence* 6, no. 1 (2021): 40-53.
- [11]. Yin, Zhixiang, Penghai Wu, Giles M. Foody, Yanlan Wu, Zihan Liu, Yun Du, and Feng Ling. "Spatiotemporal fusion of land surface temperature based on a convolutional neural network." *IEEE Transactions on Geoscience and Remote Sensing* 59, no. 2 (2020): 1808-1822.
- [12]. Zheng, Yuhui, Huihui Song, Le Sun, Zebin Wu, and Byeungwoo Jeon. "Spatiotemporal fusion of satellite images via very deep convolutional networks." *Remote Sensing* 11, no. 22 (2019): 2701.
- [13]. Jia, Duo, Changxiu Cheng, Changqing Song, Shi Shen, Lixin Ning, and Tianyuan Zhang. "A hybrid deep learning-based spatiotemporal fusion method for combining satellite images with different resolutions." *Remote Sensing* 13, no. 4 (2021): 645.
- [14]. Jia, Duo, Changxiu Cheng, Shi Shen, and Lixin Ning. "Multitask Deep Learning Framework for Spatiotemporal Fusion of NDVI." *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022): 1-13.
- [15]. Li, Jun, Yunfei Li, Runlin Cai, Lin He, Jin Chen, and Antonio Plaza. "Enhanced spatiotemporal fusion via MODIS-like images." *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021): 1-17.
- [16]. Li, Yan, Yanzhao Ren, Wanlin Gao, Jingdun Jia, Sha Tao, and Xinliang Liu. "An enhanced spatiotemporal fusion method—Implications for DNN based time-series LAI estimation by using Sentinel-2 and MODIS." *Field Crops Research* 279 (2022): 108452.
- [17]. Han, Lijing, Jianli Ding, Xiangyu Ge, Baozhong He, Jinjie Wang, Boqiang Xie, and Zipeng Zhang. "Using spatiotemporal fusion algorithms to fill in potentially absent satellite images for calculating soil salinity: A feasibility study." *International Journal of Applied Earth Observation and Geoinformation* 111 (2022): 102839.
- [18]. Bai, Bingxin, Yumin Tan, Gennadii Donchyts, Arjen Haag, and Albrecht Weerts. "A simple spatio-temporal data fusion method based on linear regression coefficient compensation." *Remote Sensing* 12, no. 23 (2020): 3900.
- [19]. Chen, Guanyu, Peng Jiao, Qing Hu, Linjie Xiao, and Zijian Ye. "SwinSTFM: Remote sensing spatiotemporal fusion using Swin transformer." *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022): 1-18.
- [20]. Li, Weisheng, Fengyan Wu, and Dongwen Cao. "Dual-Branch Remote Sensing Spatiotemporal Fusion Network Based on Selection Kernel Mechanism." *Remote Sensing* 14, no. 17 (2022): 4282.