A fast 3D convolutional neural networks based spatio-temporal fusion method using spatio -temporal-spectral dataset (STF3DCNN)

Mingyuan Peng (1) (2), Lifu Zhang (1), Xuejian Sun (1), Yi Cen (1), Xiaoyang Zhao (1)(2)

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, 20 Datun Road, Chaoyang District, Beijing, 100101, China
Aerospace and Information
[2] University of Chinese Academy of Sciences, 19 Yuquan Road, Shijingshan District, Beijing, 100049, China
Email: pengmy@radi.ac.cn; zhanglf@radi.ac.cn; sunxj@radi.ac.cn; cenyi@radi.ac.cn; zhaoxy@radi.ac.cn

KEY WORDS: spatiotemporal fusion; 3D convolution neural networks; feature learning

ABSTRACT: With the tremendous development of remote sensors, a large amount of remote sensing data is used in applications related to remote sensing, which poses new challenges to the efficiency and ability of processing big data. Spatio-temporal remote sensing data fusion can recover high spatial and high resolution remote sensing data (HSHT) from multiple remote sensing data, but the current method is time-consuming and inefficient, especially for the newly proposed deep learning-based method. Here, we propose a fast 3D convolutional neural network method based on spatio-temporal fusion using spatio-temporal spectral data set (STF3DCNN). This method can fuse low-space high-time resolution data (HTLS) and high-space low-time resolution data (HSLT) on a 4-dimensional data set of spatio-temporal spectrum, while ensuring accuracy. The method was tested on 3 data sets and ablation studies were conducted. This method is compared with the existing commonly used spatio-temporal fusion methods, which proves our conclusion.

1. Introduction

With the rapid development of remote sensing sensors and their applications, a large amount of remote sensing data has been accumulated, making it possible for applications related to long-term monitoring. Various satellites have obtained a large number of data sets with different spatial and temporal resolutions. Due to the limitations of satellite sensors, remote sensing data sets cannot have high spatial and temporal resolutions at the same time. Combining the advantages of different remote sensing products to obtain data sets with high spatial resolution and high temporal resolution has become a continuously developing research field.

Spatiotemporal data fusion is an effective choice to achieve this goal. So far, researchers have proposed many spatio-temporal fusion methods. They can be divided into 3 categories (L. Zhang, Peng, Sun, Cen, & Tong, 2019): weighted function method, linear optimization decomposition method and nonlinear optimization method. The weight function method assumes that there is a linear relationship between high-space and low-time resolution images and low-space and high-resolution images, and introduces time, space, and spectral weights into the model. The pixel center in the sliding window is used to determine the center pixel value. For example, the spatiotemporal adaptive reflection fusion model (STARFM) (Gao, Masek, Schwaller, & Hall, 2006) has been improved by scholars, the STAARCH model (Hilker et al., 2009), and the enhanced STARFM algorithm. (ESTARFM) (Zhu, Jin, Feng, Chen and Masek, 2010), mESTARFM (Fu, Chen, Wang, Zhu and Silk, 2013) and RWSTFM (J.Wang and Huang, 2017). The linear optimization decomposition method is also based on the linear assumption, and its principle is similar to the weight function method. The reconstructed image is obtained by adding constraints to obtain the best solution. Based on the optimal principle, linear optimization decomposition methods can be divided into methods based on spectral decomposition, Bayesian method and sparse representation. Commonly used algorithms include MMT spectrum decomposition algorithm (Zhukov, Oertel, Lanzl, & Reinhackel, 1999), STDFM algorithm (Mingquan, Zheng, Changyao, Chaoyang, & Li, 2012), ESTDFM (W. Zhang et al., 2013), MSTDFA (Wu, Shen, Zhang, & G? Ttsche, 2015), soft clustering (Amorós-López et al., 2013), OB-STVIUM algorithm (object-based image analysis, OBIA). Non-linear mapping method based on deep learning method that can describe nonlinear relationships well. Some scholars have also explored the use of deep learning methods for space-time fusion. In recent years, the use of convolutional neural networks (CNN) for spatiotemporal fusion algorithms, such as STFDCNN (Song, Liu, Wang, Hang, and Huang, 2018), DCSTFN (Tan, Peng, Di, and Tang, 2018), have also been proposed. The nonlinear optimization method can learn and accurately describe the nonlinear relationship between the known and missing time phase images, and has higher mobility and accuracy than the linear optimization decomposition method. However, its function is related to network architecture design and parameter settings. Poor network structure usually fails to obtain good results and requires a lot of training samples and training time.

Moreover, the increase in the number of remote data sets and application requirements has brought challenges to the processing of long-term sequence data sets, requiring methods to have the capacity to handle huge data sets. Due to the design of this algorithm, traditional methods and deep learning methods 3D convolutional networks have recently been widely used in the fields of video and computer vision. Instead of 2D ConvNet, which only performs operations in space, in 3D ConvNet, convolution and pooling operations are performed spatio-temporal. This is exactly where 3D ConvNet is very suitable for spatiotemporal feature learning (Tran, Bourdev, Fergus, Torresani and Paluri, 2015). (Height and width) and the 3 dimensions of the spectrum are divided, so it is difficult to re-divide and process the repeated sequence. Zhang et al. In 2017, the idea of a new multidimensional data set (MDD) was proposed (L. Zhang, Chen, Sun, Fu, & Tong, 2017), which can establish a spatio-temporal spectral data set and integrate 4-dimensional information. MDD can meet the correlation analysis of multiple features in different dimensions, so it is a better choice to combine with 3D convolutional neural network tools.

Based on the above reasons, we here propose a fast spatiotemporal inverse fusion method using 3D convolutional neural network tools. The data set is arranged based on the idea of MDD and executed in 4 dimensions. The architecture always uses a 3D convolutional neural network to learn the mapping between the HTLS residual sequence and the HLST data set, and adds it to the original HLST data set, thereby reconstructing the overall long-term sequence data set of HTHS. The method is easy to implement and efficient and maintains accuracy.

2. Methodology

2.1 4D data arrangements

MDD arranges the data set in the 4 dimensions of the time and space (height and width) spectrum. The storage structure of MDD is called SPAtial-Temporal-Spectral (SPATS). According to the arrangement form, it mainly consists of 5 types of data formats, namely, the time sequence in the frequency band (TSB), the time sequence in the pixel (TSP), and the time of the frequency band. Interleaving (TIB) and Time Interleaving (TIS) of the spectrum. Here, we draw on the ideas of TSB, which can be expressed as follows. The TSP storage format first stores pixel data in the first time period t 1 in the order of the first column, and then stores the data in the order of each time period t 1, and finally, according to the above rules, the time cube data is stored in chronological order.

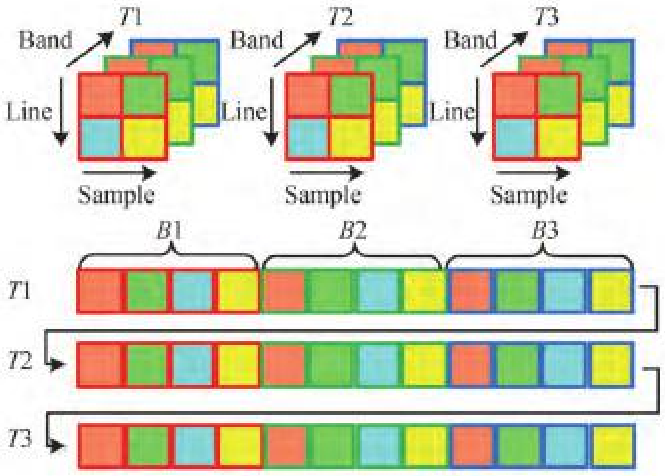


Figure 1 TSP data format of MDD

2.2 Overall Frame

The main idea of this method is to learn the mapping between the residual sequences of HTLS and HSLT. The overall architecture of this method is shown in Figure 2. It consists of two main parts: the 4D residual sequence arrangement part and the 4D residual feature mapping network. The 4D residual sequence arrangement part arranges the HSLT and HTLS data into the 4D residual sequence data set to learn 4D residual mapping and predict HTHS. The principle is shown in the illustration in 2.1. The 4D residual feature mapping network will extract features in time residuals and spatial dimensions. The architecture of the mapping network is simple and lightweight. It consists of several fully convolutional neural networks, after which the Leaky ReLu layer is set to increase nonlinearity and prevent overfitting.

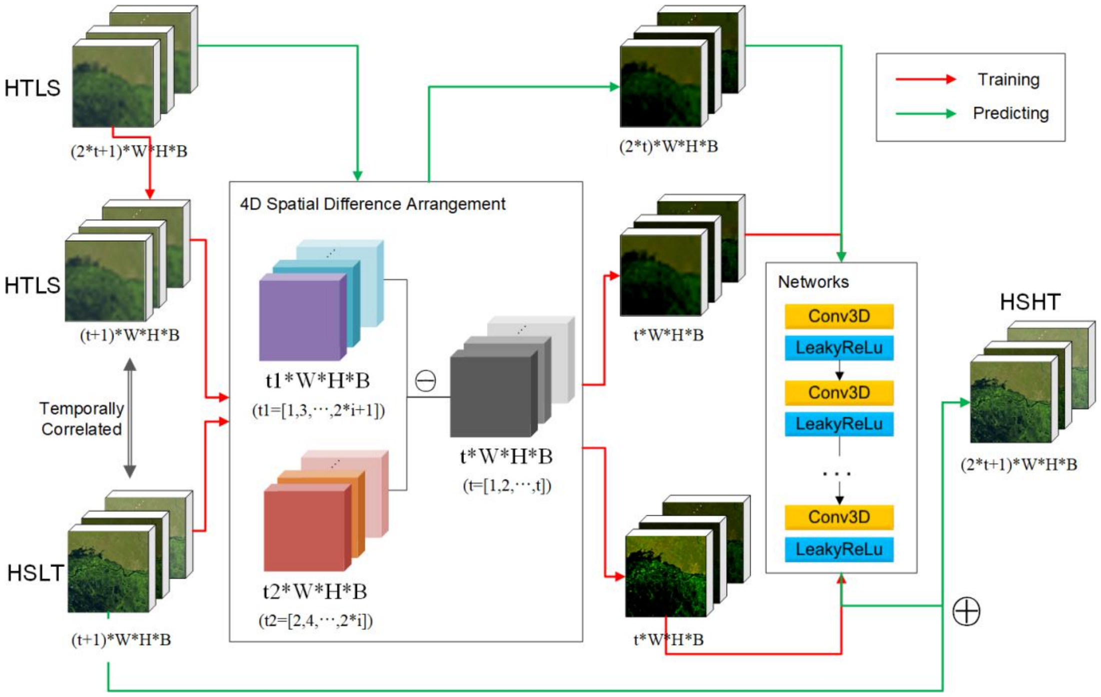


Figure 2. Overall Frame of ST3DCNN-TRS

As shown in the figure, the red line represents the training process, and the green line represents the prediction process. In the training mode, the original HTLS data set is first subsetted according to the same/close date of HSLT. First, the subset of HSLT and HLTS is used as the input of the entire training process, and then the 4D spatial residual permutation part is used for preprocessing to obtain the 4D residual sequence of HTLS (subset) and HSLT (4DRHTLS and 4RDHSLT). Then input the 4D residual sequence into the 4D residual mapping network for training. Set 4DRHTLS as network input and 4DDHSLT as network output. In the prediction mode, the original HTLS uses the 4D spatial residual permutation part for preprocessing, and obtains the 4D residual sequence of HTLS (4DRHTLS). Then input 4DDHTLS into the network to predict the simulated 4D residual sequence of HLHT (4DRHTHS). Finally, 4DRHTHS is added to the original HSLT4D, and together with the original HSLT4D, the predicted HSTH4D is generated.

3. Datasets and Experimental Settings

3.1 Datasets

In order to verify the effectiveness of this method, three data sets are used in the experiment. They are:

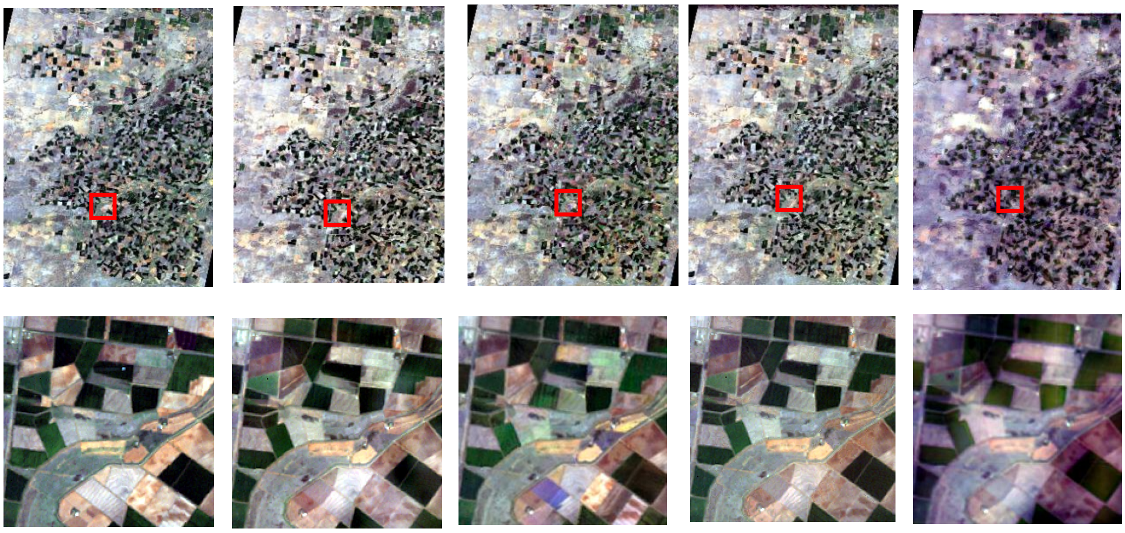
(I) Coleambally Irrigation Area (CIA) data set. It is an open source remote sensing dataset of rice irrigation system located in southern New South Wales, Australia (34.0034°E, 145.0675°S), which has been widely used in time series remote sensing research (Emelyanova, McVicar, Van Niel, Li and van Dijk, 2013 year). It contains 17 pairs of cloudless Landsat7-ETM and MODIS data detected during the summer growing season of 2001-2002. The total area is 2193 square kilometers (1720 columns of 2040 lines, with a resolution of 25 m). Initially, the nearest neighbor resampling algorithm was used to resample the MODIS data to the same spatial resolution of Landsat. But because the MODIS data contains irregularly high value points, we use the bilinear resampling algorithm to resample the MODIS data back to the original resolution and resample it back to 25m. For the sake of simplicity, all scenes are cropped into 1500 columns by 2000 rows. This data set is a good sample of seasonal changes in complex ground.

(II) The Lower Gwydir Catchment (LGC) data set. It is also an open source remote sensing dataset (Emelyanova et al., 2013) sensed in northern New South Wales (149.2815°E, 29.0855°S) from April 2004 to April 2005. It consists of 14 pairs of cloudless Landsat5 TM-MODIS, with 3200 columns by 2720 lines, and a resolution of 25 m. Similarly, the nearest neighbor resampling algorithm was initially used to resample MODIS data to the same spatial resolution of Landsat. Since MODIS data contains irregular pepper noise and is difficult to remove, the bilinear resampling algorithm is used here to synthesize MODIS data with Landsat data, and 40dB noise is added to simulate the sensing process. For the sake of simplicity, all scenes are cropped into 3000 columns by 2500 rows. On October 10, a flood occurred in this field, which made the data integrated into a good sample of a long-term sequence, and its sudden changes were more unpredictable and irregular than the CIA data set.

(III) Real Data Set (RDT). It was sensed in Louisiana (122° 39′ 34.33″ W, 39° 22′ 4.17″ N) from June to October 2013. It consists of 9 pairs of Landsat8-OLI-MODIS, and each scene is divided into 800 columns and 800 rows. Frequency band is a subset of 4 frequency bands. All data are preprocessed after geo-referencing and atmospheric correction. The bilinear resampling algorithm is used to resample the MODIS data to the same resolution as the Landsat data. This data set is a good sample of seasonal changes in crop fields and vegetated mountains.

4. Results

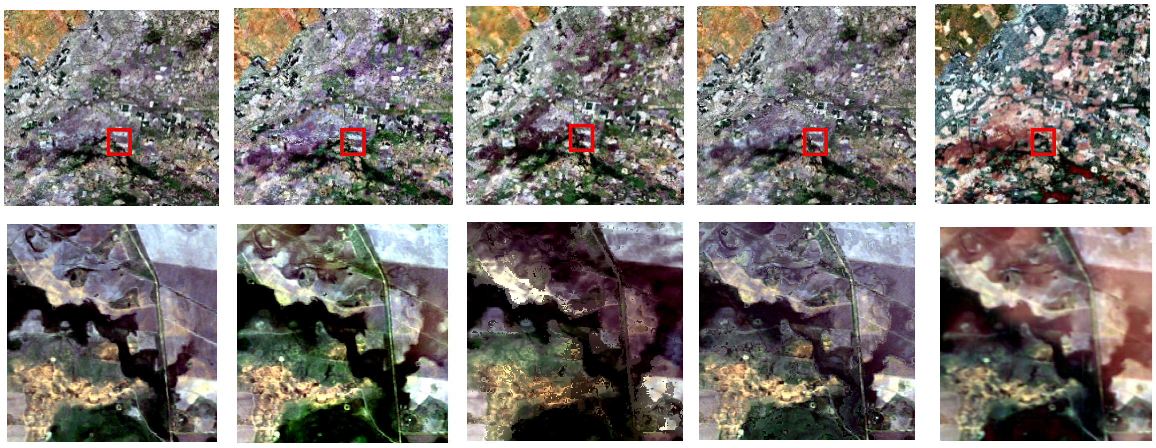
In order to compare the texture and tones of the reconstructed dataset, here we choose one date per dataset to display the fusion results in whole scene and detail. For the CIA dataset, date 10 was selected and their RGB composites are shown in the Figure 3. We can see that our method recovers the texture and tone well and similar to FSDAF. For ESTARFM some landcovers were largely mistaken. DCSTFN can recover the data yet suffer from blur. We can see that for CIA dataset, our method and FSDAF can best predict the missing image.



(a) (b) (c) (d) (e)

Figure 3. RGB composites of results of CIA at date 10(Overall and Detail)
(a. Reference image b. STF3DCNN c. ESTARFM d.FSDAF e.DCSTFN)

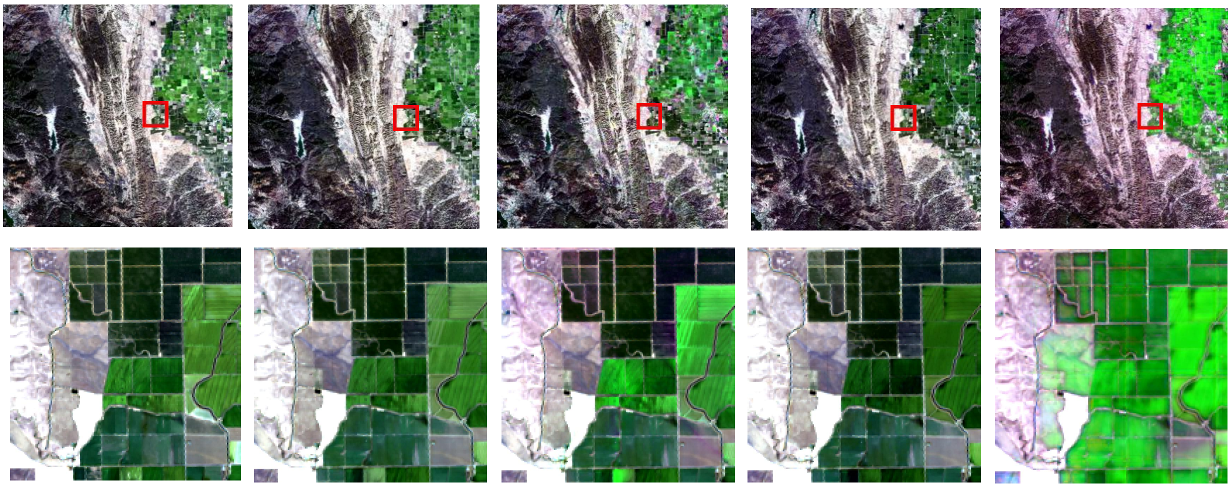
For the LGC dataset, date 10 was selected. It is the date when the flood occurs with drastic change. Their RGB composites are shown in the Figure 4. We can see from the figure that our method can recover the flooded area with accurate texture and tone. And FSDAF can also well predict the flooded area, which seemingly outperform ESTARFM. For DCSTFN the tones are a little wrong and the texture suffers from blur.



(a) (b) (c) (d) (e)

Figure 4. RGB composites of results of LGC at date 10 (Overall and Detail)
(a. Reference image b. STF3DCNN c. ESTARFM d.FSDAF e.DCSTFN)

For the RDT dataset, date 6 was selected and the RGB composites are shown in the Figure 5. The whole image and the detail(with multiple landcover types) shows that our method performs well as well as FSDAF. And for ESTARFM the tone is not completely true to the original as well as DCSTFN.



(a) (b) (c) (d) (e)

Figure 5. RGB composites of results of RDT at date 6 (Overall and Detail)
(a. Reference image b. STF3DCNN c. ESTARFM d.FSDAF e.DCSTFN)

Also, to best show the average accuracies of our methods, we calculated the average indices of each method on all dates, and record the overall running time of all time-series. The average accuracy results and total running time are shown in the tables below. We can see that though our method did not perform the best, yet it did not perform too badly. However, the total running time decreased hugely by many times. For traditional methods using CPU, our method increased the efficiency by 109 times(ESTARFM) and 104 times(FSDAF). And for the DCTSFN, which is also the deep learning spatiotemporal fusion method, the running time shrinks 12 times among three datasets. And we can find the tendency that larger the dataset, higher efficiency. The experiments proved our goal: our method maintains the average accuracy of the existing method and at the same time it hugely increase the efficiency.

Table 1. Average fusion indices results of all three datasets

	CC	SAM	RMSE	ERGAS	PSNR	Q	DD	SSIM	KGE
STF3DCNN	**0.8740**	**0.9928**	386.4355	1.3724	**33.3709**	**0.8684**	104.2675	8.51833E-06	**0.8410**
ESTARFM	0.8255	0.9892	403.6620	1.4545	29.2558	0.8158	**82.3293**	8.57584E-06	0.7364
FSDAF	0.8595	0.9920	**362.5359**	**1.3659**	29.9007	0.8525	129.0426	**8.80058E-06**	0.8216
DCSTFN	0.6969	0.9753	552.7776	2.2547	26.8396	0.6715	255.0605	5.98376E-06	0.6293

Table 2. Running times of whole time series using different method(in seconds)

	CIA/s	LGC/s	RDT/s
STF3DCNN	**552**	**987**	**77**
ESTARFM	6.40E+11	9.63E+15	14435.744
FSDAF	2.94E+06	6.40E+09	7595.211
DCSTFN	6910	12278	489.4740

5. Conclusions

In this article, we propose a fast spatiotemporal fusion method using a 3D fully convolutional neural network on a spatiotemporal spectrum dataset. Based on the idea of multidimensional data set (MDD), the long-term series data is arranged and operated in 4 dimensions according to the spatio-temporal spectrum data set. By using a 3D fully convolutional neural network to learn the mapping between the residual sequences of HTLS and HSLT, compared with existing methods, this model can efficiently restore the 4D HTHS data set and maintain accuracy.

In this article, we used 3 data sets to verify the efficiency and accuracy of the method. An ablation study was conducted to discuss the influence of parameters on long-term sequence spatio-temporal fusion under different circumstances, including network depth, whether to use time weights, and whether to use residual blocks. We found that the smaller the number of layers, the higher the accuracy and efficiency; for long corridors with seasonal and regular changes in time weight, the performance of the model is better; for land cover with sudden irregular changes, The effect of not using time weighting is better. Using residual blocks does not improve accuracy. Finally, the method is compared with existing methods in accuracy and running time. Experiments show that our method can greatly improve efficiency and maintain overall accuracy, especially when the data set is large and the time span is long.

References

Alparone, L., Wald, L., Chanussot, J., Thomas, C., Gamba, P., & Bruce, L. (2007). Comparison of Pansharpening Algorithms: Outcome of the 2006 GRS-S Data Fusion Contest. IEEE Transactions on Geoscience and Remote Sensing, 45, 3012-3021. doi:10.1109/TGRS.2007.904923

Amorós-López, J., Gómez-Chova, L., Alonso, L., Guanter, L., Zurita-Milla, R., Moreno, J., & Camps-Valls, G. (2013). Multitemporal fusion of Landsat/TM and ENVISAT/MERIS for crop monitoring. International Journal of Applied Earth Observation & Geoinformation, 23(Complete), 132-141.

Dan, L., Hao, M., Zhang, J. Q., Bo, H., & Lu, Q. (2012). A universal hypercomplex color image quality index. Conference Record IEEE Instrumentation & Measurement Technology Conference.

Emelyanova, I. V., McVicar, T. R., Van Niel, T. G., Li, L. T., & van Dijk, A. I. J. M. (2013). Assessing the accuracy of blending Landsat–MODIS surface reflectances in two landscapes with contrasting spatial and temporal dynamics: A framework for algorithm selection. Remote Sensing of Environment, 133, 193-209. doi:https://doi.org/10.1016/j.rse.2013.02.007

Fu, D., Chen, B., Wang, J., Zhu, X., & Hilker, T. (2013). An Improved Image Fusion Approach Based on Enhanced Spatial and Temporal the Adaptive Reflectance Fusion Model. Remote Sensing, 5, 6346-6360. doi:10.3390/rs5126346

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

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition.

Hilker, T., Wulder, M. A., Coops, N. C., Linke, J., Mcdermid, G., Masek, J. G., . . . White, J. C. (2009). 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.

Huynh-Thu, Q., & Ghanbari, M. (2008). Scope of validity of PSNR in image/video quality assessment. Electronics Letters, 44(13), 800-801.

Kingma, D., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. Computer ence.

Lu, M., Chen, J., Tang, H., Rao, Y., Yang, P., & Wu, W. (2016). Land cover change detection by integrating object-based data blending model of Landsat and MODIS. Remote Sensing of Environment, 184, 374-386.

Mingquan, W., Zheng, N., Changyao, W., Chaoyang, W., & Li, W. (2012). 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), 1-14. doi:10.1117/1.JRS.6.063507

Moosavi, V., Talebi, A., Mokhtari, M. H., Shamsi, S. R. F., & Niazi, Y. (2015). A wavelet-artificial intelligence fusion approach (WAIFA) for blending Landsat and MODIS surface temperature. Remote Sensing of Environment.

Song, H., Liu, Q., Wang, G., Hang, R., & Huang, B. (2018). Spatiotemporal Satellite Image Fusion Using Deep Convolutional Neural Networks. IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing, 1-9.

Tan, Z., Peng, Y., Di, L., & Tang, J. (2018). Deriving High Spatiotemporal Remote Sensing Images Using Deep Convolutional Network. Remote Sensing, 10(7), 1066-.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015, 7-13 Dec. 2015). Learning Spatiotemporal Features with 3D Convolutional Networks. Paper presented at the 2015 IEEE International Conference on Computer Vision (ICCV).

Wang, J., & Huang, B. (2017). A Rigorously-Weighted Spatiotemporal Fusion Model with Uncertainty Analysis. Remote Sensing, 9, 990. doi:10.3390/rs9100990

Wang, Z. (2004). Image Quality Assessment : From Error Visibility to Structural Similarity. IEEE Transactions on Image Processing.

Wei, Q., Dobigeon, N., & Tourneret, J. Y. (2015). Bayesian fusion of multispectral and hyperspectral images with unknown sensor spectral response. 2014 IEEE International Conference on Image Processing, ICIP 2014, 698-702. doi:10.1109/ICIP.2014.7025140

Wu, P., Shen, H., Zhang, L., & G?Ttsche, F. M. (2015). Integrated fusion of multi-scale polar-orbiting and geostationary satellite observations for the mapping of high spatial and temporal resolution land surface temperature. Remote Sensing of Environment, 156, 169-181.

Zhang, L., Chen, H., Sun, X., Fu, D., & Tong, Q. (2017). Designing spatial-temporal-spectral integrated storage structure of multi-dimensional remote sensing images. Yaogan Xuebao/Journal of Remote Sensing, 21, 62-73. doi:10.11834/jrs.20176091

Zhang, L., Peng, M., Sun, X., Cen, Y., & Tong, Q. (2019). Progress and bibliometric analysis of remote sensing data fusion methods (1992—2018). Journal of Remote Sensing, 23(4), 1993-2002. doi:10.11834/jrs.20199073

Zhang, W., Li, A., Jin, H., Bian, J., Zhengjian, Z., Lei, G., . . . Huang, C. (2013). An Enhanced Spatial and Temporal Data Fusion Model for Fusing Landsat and MODIS Surface Reflectance to Generate High Temporal Landsat-Like Data. Remote Sensing, 5, 5346-5368. doi:10.3390/rs5105346

Zhang, Y., De Backer, S., & Scheunders, P. (2009). Noise-Resistant Wavelet-Based Bayesian Fusion of Multispectral and Hyperspectral Images. IEEE Transactions on Geoscience & Remote Sensing, 47(11), p.3834-3843.

Zhu, X., Jin, C., Feng, G., Chen, X., & Masek, J. G. (2010). An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions. Remote Sensing of Environment, 114(11), 2610-2623.

Zhukov, B., Oertel, D., Lanzl, F., & Reinhackel, G. (1999). Unmixing-based multisensor multiresolution image fusion. IEEE Transactions on Geoscience & Remote Sensing, 37(3), 1212-1226.

Zurita-Milla, R., Clevers, J. G. P. W., & Schaepman, M. E. (2008). Unmixing-based landsat TM and MERIS FR data fusion. Geoscience and Remote Sensing Letters, IEEE, 5, 453-457. doi:10.1109/LGRS.2008.919685