Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 8, No. 1, 2023

Improved GSA Method Based on Deep Convolutional Features

Shuai Wu

School of Surveying and Land Information Engineering, Henan Polytechnic University, Jiaozuo 454000, China

Abstract: With the rapid development of remote sensing platforms, sensors and other technologies, remote sensing, as an important technical means of collecting information on land cover and its changes, plays an important role in land cover classification and dynamic monitoring. Due to the limitation of the imaging mechanism of in-orbit optical remote sensing sensors, it is difficult for to take into account both spatial resolution and spectral resolution. There is a need for complementary use of different sensor images and to improve the accuracy of feature interpretation. In this study, based on the Gram-Schmidt Adaptive (GSA) and its classical extension, a neural network-based multi-source remote sensing image fusion method is proposed by combining the U-Net coding and decoding structure with the Non-Local spatial attention mechanism. Using two sets of experimental datasets, three kinds of fusion evaluation metrics without reference and three kinds of fusion evaluation metrics with reference, the improved method and the existing fusion method are compared.

Keywords: Multi-source remote sensing imagery; GSA; U-net.

1. Introduction

With the rapid development of remote sensing platforms, sensors, communications and data processing technologies, remote sensing technology, as an important means of information collection, has been able to achieve all-weather, all-day, all-round observation of the Earth's data and information on its changes[1,2] . Meanwhile, in the 1980s, the introduction of hyperspectral earth observation technology, which combines the spectra of features with their spatial and geometric information, marked the entry of remote sensing technology into a completely new stage of development[3] . The high spectral resolution, spectral integration and spectral continuity of spectral remote sensing images have opened up new ways for image analysis and information extraction[4,5] . In addition, the emergence of high-resolution images can provide rich geometric structure, spatial features, shape texture and other detailed information of features, which realises the fine description of surface targets[6,7] . Given the significant constraints between the spatial and spectral resolutions of remote sensing sensors, the acquisition of High Resolution Multispectral (HRM) imagery using existing equipment can be quite challenging[8] . Therefore, multi-source data fusion is needed to organically combine the information contained in the images from different sensors in a complementary manner, spatially enhance the lower resolution spectral data, or equivalently provide better spectral resolution for the higher resolution data, in order to improve the reliability of image interpretation, the ability to interpret the images, and thus improve the accuracy of data classification and target identification [9,10] .

Multi-source optical remote sensing image fusion technology is a technology that fuses remote sensing images collected by a variety of different resolutions and sensors. It makes full use of the advantages of different remote sensing sensors, fuses remote sensing image information from different sources, such as high spatial resolution, high spectral resolution, high temporal resolution, etc., and fuses remote sensing images from different sources and resolutions into an image with high spatial and spectral resolution[11] . The integrated use of these data can obtain more surface coverage information, such as surface temperature, vegetation cover, water distribution, etc., to enhance the completeness and reliability of the data source, so as to better extract the feature information, provide a solid foundation for the subsequent applications, and improve the accuracy and precision of the feature identification and the accuracy of the target detection and classification [12-14] .

2. Experimental Data and Accuracy Evaluation Method

2.1. Experimental data

The experimental study area of this paper is located in the northwestern part of Beijing, China. Beijing is located in the northern part of China and has a warm-temperate semi-humid climate, semi-arid climate with rainy and high temperatures in summer, cold and dry winters, and short springs and autumns. As the high-resolution data are not easy to obtain, the high-resolution data and Sentinel-2 data used in this study contain two different feature categories for urban and suburban areas to satisfy the comparative analyses.

The multispectral images with sub-metre spatial resolution of Gaofen-2 are used as high-resolution multispectral HRMS images, and the Sentinel-2 images with 10m spatial resolution in 9 bands synthesised from blue, green, red, red-edge, near-infrared, and short-wave infrared bands are selected as low-resolution multispectral LRMS images.

The same area contains one view of GF-2 data and one view of Sentinel-2 data, and the GF-2 data and Sentinel-2 data were ingested close to each other, but with a few days difference in time. The urban data have many and dense buildings, which are highly susceptible to shadows on the image, and the roads are mainly tarmac with a small amount of bare ground in the upper right corner of the image. The suburban data are mostly bare land, arable land and industrial land with a large number of water bodies sporadically distributed. The fusion raw data containing different rich features are selected to better analyse the robustness of the fusion method in different scenarios. The true colours of the

study area are shown in Fig. 1, where Figs. (a) and (b) correspond to the urban area data, which represent the high-resolution GF-2 data and the multispectral Sentinel-2 data, respectively; Figs. (c) and (d) correspond to the suburban area data. The specific information of the experimental data is shown in Table 1.

Table 1. Basic information of experimental data

Area	Satellite	Pixel size	Date	Num
Urban	GF-2	3000*3000	2017/12/13	(a)
	Sentinel-2 (S2)	240*240	2017/12/24	(b)
Suburban	GF-2	3000*3000	2017/12/13	(c)
	Sentinel-2 (S2)	240*240	2017/12/24	(d)

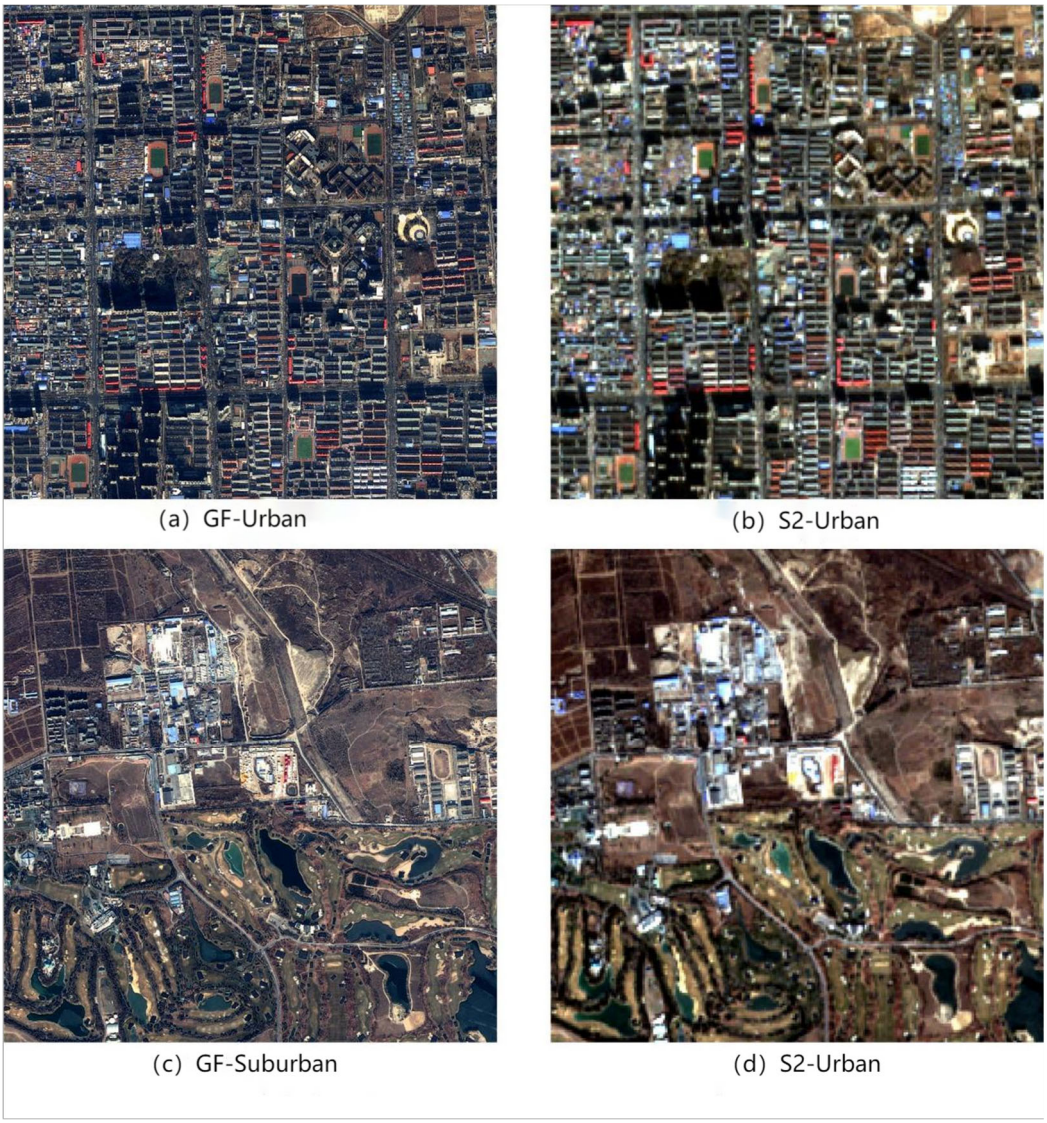

Figure 1. True colour image in the study area

2.2. Accuracy evaluation method

Classical image fusion evaluation methods are mainly divided into two kinds: subjective evaluation and objective evaluation, respectively, from the visual and index numerical examination of the fusion results of the advantages and disadvantages of the two evaluation modes complementary integrated measure of the resultant image quality.

(1) Subjective evaluation methods for multi-source image fusion

The subjective evaluation of the fused image usually selects part of the fusion result to be displayed in true colour or pseudo-colour, and then the human eye is used to compare and observe the fusion effect of different methods or different parameters, and the observer evaluates the subjective evaluation through the aspects of whether the image features are distinctive or not, whether the texture is clear or not, whether the image is distorted or not, and whether the degree of loss of image information is lost or not, etc. The advantage of this method lies in the fact that the observer can give the good or bad image fusion quality very quickly and easily. The advantage is that the observer can easily and quickly give a good or bad image fusion quality, which is very intuitive in the evaluation of the quality of near-infrared and visible image fusion. This evaluation method is more intuitive and does not rely on numerical calculations, but it requires that the observer needs to have enough experience in order to give a precise evaluation, and it is not applicable to large-scale image processing situations because of the low efficiency of

the human eye's visual comparison. In addition, since the human eye visual system judgement is not very stable, this also affects the accuracy of subjective visual evaluation from time to time.

(2) Objective evaluation index for multi-source image fusion

Although the subjective evaluation method is more intuitive, it is subject to the interference of human subjective factors, and it is difficult for the observer to grasp the evaluation standard, so the subjective evaluation is one-sided and unstable. The objective evaluation method is different from the subjective evaluation method, which can reflect the overall quality of the image fusion through the statistical value of the relevant indicators, can overcome the interference of the visual factors of the observers, and has the advantages of quantitative and repeatable. Therefore, the qualitative analysis of subjective evaluation and the quantitative analysis of objective evaluation should be combined to construct the image fusion evaluation index.

The objective evaluation methods are generally divided into those that do not require reference images and those that require reference images[15] , for the former, three indexes of band grey mean, information entropy and mean gradient are used, and for the latter, three indexes of spectral angle SAM, relative overall dimensionless error ERGAS and general image quality index Q4 are selected, as shown in Table 2.

Among them, the mean grey scale reflects the overall brightness of the image, if the fusion result has a large difference with the average value of the reference image, it indicates that the method introduces a large low-frequency signal distortion. The information entropy reflects the information increase of the fused image, the larger the entropy, the larger the information, and the richer the detail information; the average gradient reflects the spatial detail integration, which can also reflect the overall level of the image edge strength, the larger the average gradient, the stronger the sharpening effect of the image, and the higher the local contrast; the SAM and the ERGAS are the more common response to the spectral retention ability of the index; the closer the SAM is to 0, the stronger the spectral retention ability is; the closer the ERGAS is to 0, the stronger the spectral retention ability is. The closer the SAM value is to 0, the stronger the spectral retention ability is, and the closer the ERGAS value is to 0, the higher the spectral quality of the fused image is. $Q4$ The closer the value is to 1, the higher the image quality.

Table 2. Fusion evaluation indicators

Evaluation indicators	Expression formula	physical meaning
gamma-mean (statistics)	$mean = \frac{1}{N}\sum_{i=1}^{N} x_i$	It is important to reveal the error characteristics of individual bands because they have a significant impact on many applications based on spectral indices and band ratios. indices and band ratios
entropy	$Entropy = -\sum_{i}^{L} P(i)\lg P(i)$	Entropy is mainly a measure of how much information is contained in the image, and the higher the information entropy, the richer and better the quality of the fused image.
mean gradient	$G = \frac{1}{M \times N}\sum_{i=1}^{M}\sum_{j=1}^{N}\sqrt{\frac{\left(\frac{\partial f}{\partial \mathrm{x}}\right)^2 + \left(\frac{\partial f}{\partial \mathrm{y}}\right)^2}{2}}$	The average gradient reflects the detail contrast and texture transformation in the image, it reflects the clarity of the image to some extent.
SAM	$SAM = \frac{1}{N}\sum_{i=1}^{N}\cos^{-1}\frac{x_i \cdot y_i}{\lvert x_i \rvert \cdot \lvert y_i \rvert}$	Reflects the similarity between spectra, the smaller the angle the higher the similarity and vice versa.
$ERGAS$	$ERGAS = 100\frac{H}{L}\sqrt{\frac{1}{n}\sum_{i=1}^{n}[RMSE(i)/Mean(i)]^2}$	The closer ERGAS is to 0, the higher the spectral quality of the fused image.
$Q4$	$Q(R,F) = \frac{4\sigma_{RF}\cdot\mu(R)\cdot\mu(F)}{\left(\delta_R^2+\delta_F^2\right)\left(\mu_R^2+\mu_F^2\right)}$	A sliding window is used to increase the discriminatory power and to measure the local distortion of the fused image, and all the sliding window is averaged .

3. Improved GSA method based on deep convolutional features

3.1. Convolutional neural network

Convolutional Neural Network (CNN) is a feed-forward neural network that has the ability to exploit spatial correlations in data to effectively capture intrinsic relationships in raster data, and has features such as automatic feature extraction, weight sharing and hierarchical learning, which are useful in computer vision tasks such as image segmentation, classification, monitoring[16] and image fusion, etc. Therefore, the aim of this study is to extract the null spectral features of images using convolutional neural network structure.

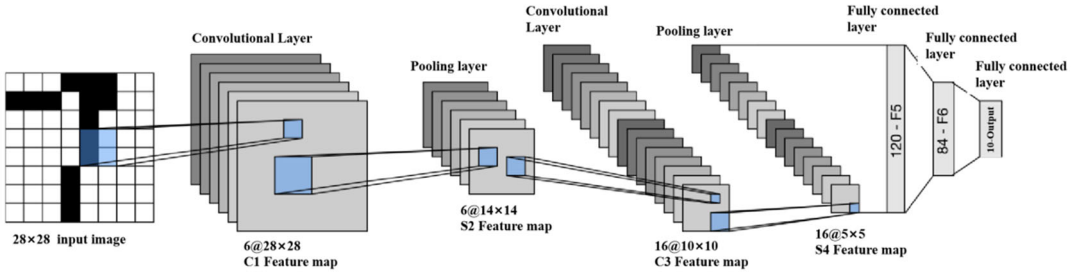

Figure 2. Structure diagram of convolution neural network[17]

The classical CNN structure basically consists of a combination of structures such as convolutional layers, pooling layers, activation layers, etc., and covers one or more fully connected layers on the tail end, the convolutional neural network structure is shown in Figure 2.

The convolutional layer consists of a set of convolutional kernels, usually the size of the convolutional kernel is smaller than the size of the input channel, so it is necessary to select a certain range of the input channel as the convolutional window. The convolution window represents the range of the image perceived by the convolution kernel and is usually the same size as the convolution kernel. Presented on the image, the convolution window moves from left to right and from top to bottom, and after each move, the convolution kernel is multiplied with the corresponding image slice at the current position to extract the features at that position. The convolution process can be specifically described as follows:

$$Y_{i,j} = \sum_{i=0}^{m} \sum_{j=0}^{n} W_{m,n} \cdot X + W_b \quad (1)$$

where$Y_{i,j}$ is the result of sliding the image(i, j) after the convolution of the convolution window at$W_{m,n}$ which is the convolution kernel tensor of size m * n, and$X$ is the tensor of image slices in the current convolution window, andW_b is the bias matrix. The size of the convolution kernel tensor is generally set to 3*3, 5*5 or 7*7, etc. The larger the convolution kernel is, the more parameters the model has, the stronger the ability to extract the deeper semantics, but the higher the demand for computing power. Therefore, it is proposed to replace the large convolutional kernel with a stack of multiple small convolutional kernels with the same sense field (e.g., the effect of stacking two 3*3 convolutional kernels is equal to that of 5*5 convolutional kernels), and to reduce the number of parameters without reducing the ability to extract the semantics through stacking. The stacking approach reduces the number of parameters without reducing the semantic extraction capability.

3.2. U-Net codec structure

U-Net network is a network model for medical image segmentation proposed by Ronneberger et al. in 2015, which is widely used in the field of image segmentation. The model is divided into two parts: encoder and decoder. The encoder part is responsible for multi-scale feature extraction, and adopts the classical convolutional neural network structure such as VGG16 or VGG19, which obtains images at different scales by downsampling the images several times, and then convolves the images at different scales to obtain features at different levels. The decoder part adopts tansposed convolution and skip connection, which is responsible for fusing the extracted features and generating the segmentation results, up-sampling the features at different scales to recover the position information of the features, fusing the features at different scales by splicing, and finally obtaining the segmentation results after multiple convolutions. Segmentation results [18] Segmentation results are obtained after multiple convolutions.

The two main advantages of U-Net are (1) the skip connection approach can be used by passing features from different stages of the encoder to the decoder, providing a large amount of spatial and detailed information for resolution reconstruction, which results in a higher quality reconstruction; (2) progressive resolution reconstruction, which is different from the early full convolutional networks, and this approach reduces the reconstruction step size, which is conducive to the model's self correction also reduces the learning difficulty. This ensures that the final recovered feature map incorporates more low-level features, and also allows the fusion of features at different scales, thus allowing for multi-scale prediction and deep-level supervision. Meanwhile, the 4 times up-sampling also makes the information such as image edges more fine.

The unique "U" structure of U-Net makes it possible to obtain the deep spatial information of the image through a small amount of computation, which makes the details of the image prediction more accurate, so this paper adopts a similar coder-decoder structure as the architecture of image fusion.

3.3. Non-Local Spatial Attention

The convolution unit in CNN only focuses on the region of the neighbourhood kernel size each time, even if the feeling field is getting bigger and bigger in the later stage, it is still the operation of the local region after all, which ignores the contribution of the other slices of the global (e.g., very far pixels) to the current region. Xiaolong Wang et al.[19] proposed a self-attention model Non-Local Neural Network in 2018, which is somewhat similar to Non-Local Means non-local mean-desiccation filtering. Ordinary filtering involves a 3×3 convolutional kernel, which is then shifted over the entire image, dealing with information that is localised to the 3×3. The Non-Local Means operation, on the other hand, combines a relatively large search range and weights it. Similarly, Non-Local is similar to the above operations, but it focuses on the extent of the receptive field in the neural network. Generally convolutional receptive fields are 3×3 or 5×5 in size, whereas using Non-Local allows the receptive field to be very large rather than being confined to a local field.Non-Local captures the relationship of the remote, which for multi-channel remote sensing images is counting all the pixels in all the channels, and weighting the relationship of the current pixel.

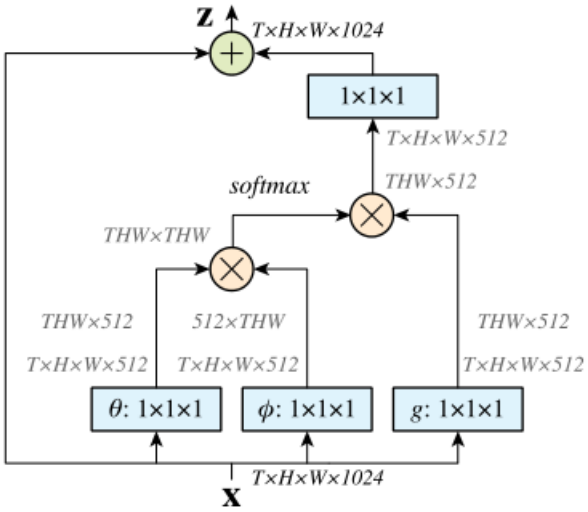

Figure 3. Non-Local Spatial Attention Structure[82]

The structure of Non-Local is shown in Figure 3.Non-Local can be used as a flexible component with a convolutional or recursive layer. It can be added to the early parts of a deep neural network, unlike the fully connected layers that are often used eventually. This allows the model to build richer hierarchies that combine non-local and local information.

3.4. Multi-source fusion network construction based on U-net and Non-Local spatial attention

In order to better fuse multi-source data to construct synthetic components, and then reconstruct the spatial spectrum features of multi-source data. In this paper, based on the GSA fusion method, a multi-source data fusion network is constructed by using the U-Net coder-decoder structure combined with the Non-Local spatial attention mechanism, and its specific structure is shown in Figure 4.

The degree of correlation between different channels and the final prediction result is first modelled through the attention mechanism, so that the network can correct the weights of different channels by learning loss, and then the data of different channels are weighted and input into the encoder. The multi-source remote sensing data are first consecutively subjected to three 3*3 convolutions in the encoder to obtain spatial features at four different scales including the original data scale, respectively, which are adopted after which the data at each scale are used to learn the remote relationship at the current scale through a Non-Local module.

Finally the model aggregates the null spectrum features of different channels by a 1*1 convolution and outputs the fusion result.

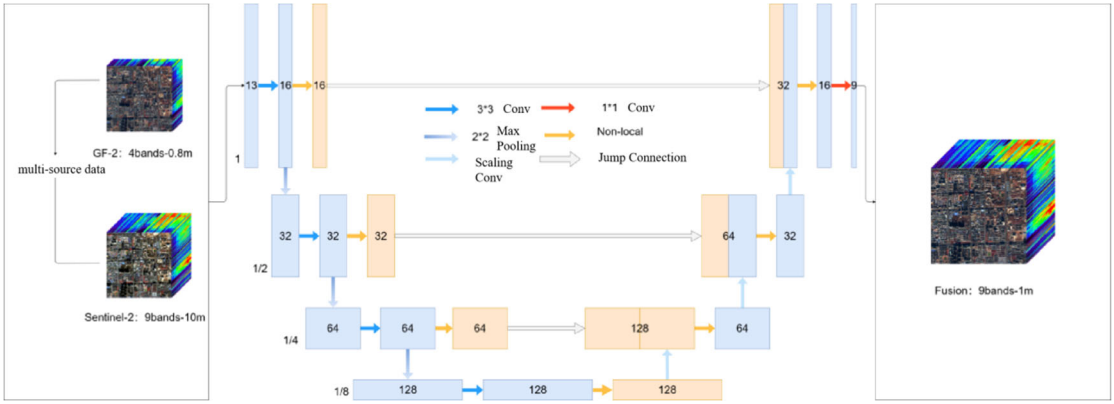

Figure 4. Structural diagram of multi-source data fusion network based on spatial attention and U-net coder-decoder

4. Analysis of Integration Results

The experiments in this chapter analyse the results of the fusion experiments of GF-2 images and Sentinel-2 images based on the close proximity of winter time in the urban and suburban areas. In order to better analyse the fusion effect of the improved fusion methods, a total of four fusion result images of the GSA-NET fusion methods and the original Sentinel-2 images are analysed from subjective qualitative and objective quantitative aspects, respectively, to prefer the best fusion method.

Firstly, from the visual effect, compared with the original multispectral image, the four fusion images have greatly improved the spatial resolution of the image. Firstly, from the visual effect, compared with the original multispectral image, all seven fusion images greatly improve the spatial resolution of the image.The PCA fusion result is overall dark, and the spectral aberration is more serious, but the clarity of the image as well as the integration of texture details is better.The GIHS fusion result retains the spectral information well, but the overall clarity is not high, and the roads, buildings and water bodies in the image have a certain degree of blurring GSA-

FCM will still have the same spatial details as this paper's method in terms of better retention of spectral information while the incorporation of spatial texture details is also greatly enhanced.

The fusion results of GSA-NET are shown in Table 3 and Figure 5:

Table 3. GSA-NET fusion accuracy evaluation

GSA-NET	wave band	average value	information entropy	mean gradient	Sam.	ERGAS	Q4
Urban	B	763.471	2.256	2.653	3.256	2.301	0.669
	G	858.562	2.352				
	R	1020.513	6.103				
	Red-Edge	1188.441	6.384				
	NIR	1298.411	6.937				
	SWIR	1212.525	6.128				
Suburban	B	721.675	2.036	2.517	2.930	1.551	0.766
	G	946.406	2.144				
	R	1208.217	6.039				
	Red-Edge	1707.968	5.992				
	NIR	2088.056	7.123				
	SWIR	1827.983	5.496				

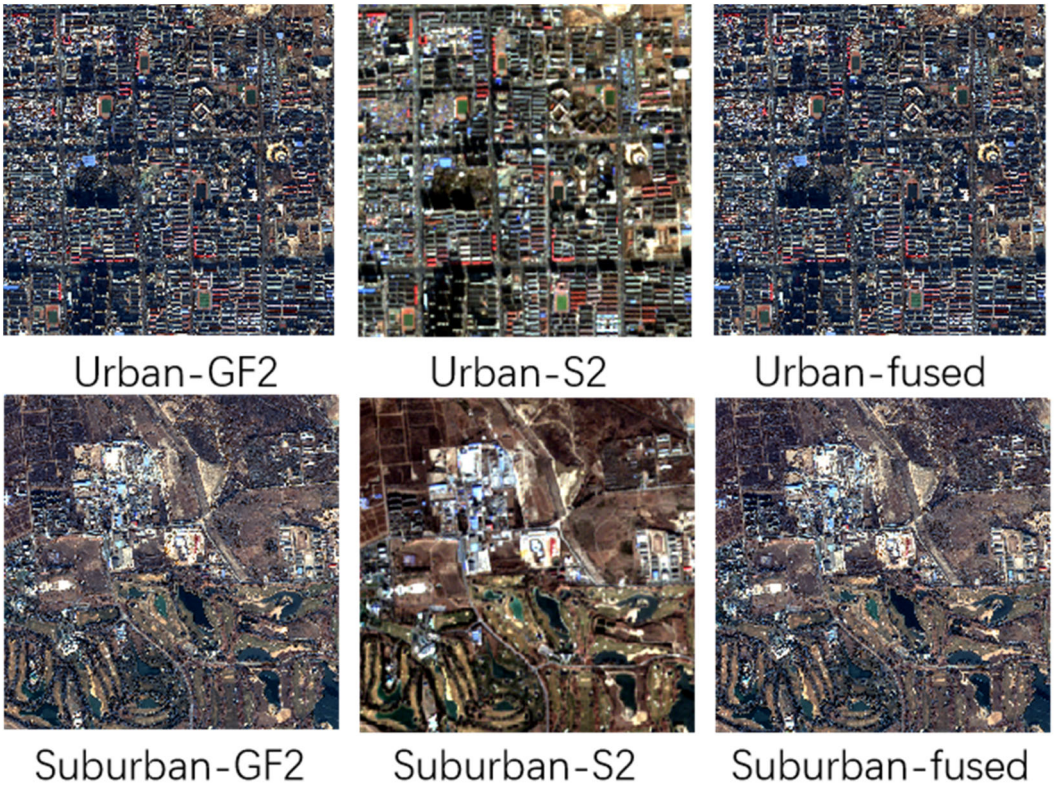

Figure 5. GSA-NET fusion of original and result images

5. Conclusion

In this paper, using GF-2 and Sentinel-2 data, we propose an improved GSA fusion formula, GSA-NET, to address the problems of the traditional CS component substitution fusion in multi-source data fusion, and verify its fusion performance and stability on different datasets by using the subjective and objective evaluation methods and classification accuracy.

To study the fusion performance of the GSA improved method, this paper selects two typical sample areas in the town as the study area, and evaluates the results using three reference-free fusion evaluation indexes and three reference fusion evaluation indexes, respectively, and compares them with the results of the other fusion methods and the reference images to prove the advantages of the proposed method. The experimental conclusions show that compared with other methods, the GSA-NET method can still maintain good spectral information while highly incorporating details, has the largest increment of information entropy, and the fused image contains the richest information with better comprehensive performance.

From the final results, the experiments in this paper have achieved the research purpose, and the proposed GSA-NET method performs well compared with the traditional classical algorithms in multi-source data fusion applications, but due to the limited time and level, there are still some deficiencies, which need to be further studied carefully. In the future, hybrid methods can be developed to combine the advantages of different categories of methods, for example, introducing particle swarm algorithms and swarm intelligence algorithms to improve the traditional fusion methods.

References

[1] ZHANG Bing, GAO Lianru, LI Jiaxin, et al. Progress and prospect of super-resolution fusion research on high/multispectral remote sensing images[J]. Journal of Surveying and Mapping, 2023, 52(7):1074-1089.

[2] Zhang Y , Ji Q . Active and dynamic information fusion for facial expression understanding from image sequences[J]. IEEE Trans Pattern Anal Mach Intell, 2005, 27(5):699-714.

[3] TONG Qingxi, ZHANG Bing, ZHANG Lifu. Frontier progress of hyperspectral remote sensing in China[J]. Journal of Remote Sensing, 2016, 20(5): 689-707.

[4] Nencini F , Garzelli A , Baronti S , et al. Remote sensing image fusion using the curvelet transform[J]. Information Fusion, 2007, 8(2):143-156.

[5] Aiazzi B , Alparone L , Baronti S ,et al. 25 years of pansharpening: a critical review and new developments[M]. 2012.

[6] Vivone G , Alparone L , Chanussot J , et al. A Critical Comparison Among Pansharpening Algorithms[J]. IEEE Transactions on Geoence & Remote Sensing, 2015, 53 (5): 2565-2586.

[7] Zhang Liangpei, Shen Huanfeng. Progress and foresight of remote sensing data fusion[J]. Journal of Remote Sensing, 2016, 20(5):12.

[8] LI Shutao,LI Congyu,KANG Xudong. Current status and future outlook of the development of multi-source remote sensing image fusion[J]. Journal of Remote Sensing, 2021, 25(1): 148-166.

[9] Vivone G , Restaino R , Licciardi G A , et al. MultiResolution Analysis and Component Substitution techniques for hyperspectral Pansharpening[C]// Geoscience & Remote Sensing Symposium. ieee, 2014.

[10] Ghassemian, Hassan. a review of remote sensing image fusion methods[J]. Information Fusion, 2016:75-89.

[11] Yokoya N , Grohnfeldt C , Chanussot J . Hyperspectral and Multispectral Data Fusion: a comparative review of the recent literature[J]. IEEE Geoscience & Remote Sensing Magazine, 2017, 5(2):29-56.

[12] Yan Junhua, Zhang Kun, Shi Tianjun, et al. Detection of weak targets on the ground of remote sensing images by fusing multilevel features[J]. Journal of Instrumentation, 2022 (003): 043.

[13] XU Shengjun, ZHANG Ruoxuan, MENG Yuebo,et al. Fusion of fractal geometric features for building segmentation in Resnet remote sensing images [J]. Optical Precision Engineering, 2022, 30(16):15.

[14] Ghamisi P , Rasti B , Yokoya N , et al. Multisource and Multitemporal Data Fusion in Remote Sensing: a Comprehensive Review of the State of the Art[J]. IEEE Geoscience and Remote Sensing Magazine, 2019, 7(1):6-39.

[15] Vivone G, Alparone L, Garzelli L, and Lolli S. Fast reproducible pansharpening based on instrument and acquisition modelling: AWLP revisited[J]. Remote Sensing, 2019, 11(19):2315:1-2315:23.

[16] Shi Y, Tian Y, Wang Y, et al. Sequential Deep Trajectory Descriptor for Action Recognition With Three-Stream CNN[J]. IEEE Transactions on Multimedia, 2017, 19(7): 1510-1520.

[17] Tian C, Zhuge R, Wu Z, et al. Lightweight image super-resolution with enhanced CNN[J]. Knowledge-Based Systems, 2020, 205: 106235.

[18] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]//Navab N, Hornegger J, Wells W M, et al. Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015. cham: Springer International Publishing, 2015: 234-241.

[19] Wang X, Girshick R, Gupta A, et al. Non-local neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. 7794-7803.