

Fusion and Super-Resolution of Remote Sensing Images Using Deep Learning and Gram-Schmidt Orthogonalization

Sin-Gil Kim^{1*}, Il-Kwang Jo², Gyong-Song Chae², Ji-Chon Ro²

¹Department of Information Science, University of Sciences, Unjong District, Pyongyang, Democratic People's Republic of Korea

²Institute of Earth Environmental Geoinformatics, Unjong District, Pyongyang, Democratic People's Republic of Korea

*Correspondence: ksg1957@star-co.net.kp

Received: 30 May 2025 / Accepted: 26 February 2026 / Published: 05 March 2026

Abstract: Multispectral satellite images contain rich spectral information but often have low spatial resolution. Super-resolution of these images is important for resource development, environmental protection, disaster prevention, regional ecology, and change detection. Deep learning models, including convolutional neural networks (CNNs), have been widely applied to the super-resolution of satellite images, including Sentinel-2 images. However, their output cannot exceed the maximum spatial resolution of the sensor. Here, we propose a method that first enhances the 20 m Sentinel-2 bands to 10 m using a CNN and then fuses them with 1 m aerial images by Gram-Schmidt orthogonal transformation and its inverse, while preserving multispectral information. The method was tested on the fusion of aerial and Sentinel-2 images to visualize vegetation distribution in urban areas. The proposed network achieved a root mean square error (RMSE) of 43.5 and a correlation coefficient (CC) of 0.941, compared with 123.4 and 0.811 for bicubic interpolation and 117.8 and 0.866 for area-to-point regression kriging (ATPRK). The fused images appeared as sharp as the aerial images while preserving the shape of the original spectral curves.

Keywords: Sentinel-2, pan-sharpening, convolutional neural network, aerial imagery, urban vegetation

INTRODUCTION

The fusion and super-resolution of remote sensing images is of great significance in expanding the availability of different satellite image data used to investigate and analyze natural environmental conditions for purposes such as resource development, global environmental protection, disaster prevention, regional ecology, and global change detection (Xiong et al., 2020). The widely used Sentinel-2 satellite images contain valuable information covering all continents, climatic zones, biomes, and land cover types (Lanaras et al., 2018; Drusch et al., 2012). The Sentinel-2 images contain 13 spectral bands from the visible to the shortwave infrared, with spatial resolutions of 10 m, 20 m, and 60 m, and also significant information (e.g., on vegetation) that is not included in aerial images, which have a much higher spatial resolution. Therefore, Sentinel-2 images are widely used for satellite image analysis because of this information.

However, due to the limitations of existing sensor technologies, optical sensors do not directly capture high-resolution multispectral images, and generally detect multispectral images with low spatial resolution and panchromatic images with high spatial resolution. For practical applications, multispectral images with high spatial resolution are necessary. Hence, super-resolution techniques such as panchromatic sharpening (pan-sharpening) have been proposed to obtain high-resolution multispectral images by combining the information of panchromatic and multispectral images. The panchromatic sharpening method, a typical image fusion method, fuses a low-resolution multispectral image with a high-resolution panchromatic image to produce a multispectral image with the high spatial resolution of the panchromatic band.

For fusion and super-resolution of remote sensing images, component substitution, multi-resolution analysis, and model-based methods have traditionally been used. The component substitution upsamples multispectral images to the same resolution as panchromatic images, transforms these images into new feature spaces, replaces components that reflect spatial information in multispectral images with panchromatic images, and performs the inverse transformation of their replaced multispectral images to obtain the final fusion image. Examples of component substitution are IHS (Intensity-Hue-Saturation),

PCA (Principal Component Analysis), GHS (Generalized Intensity Hue Saturation), Brovey transform, and Gram-Schmidt method (Aiazzi et al., 2007; Vivone et al., 2015). The component substitution is simple, easy to implement, and has the advantage of greatly improving the spatial information of multispectral images, but it distorts the spectral information of the fusion results. The multi-resolution analysis is a method of injecting high-frequency information of panchromatic images into multispectral images that are upsampled, using the wavelet, curvelet, contourlet, and Laplacian pyramid transforms. The multi-resolution analysis is superior to component substitution in preserving spectral information, but has the disadvantage of distorting spatial information. The model-based approach is based on mathematical models that describe the relationship between low-resolution multispectral images and high-resolution multispectral images, which include Markov random fields, variational methods, and sparse representations. However, the mapping relationship between low-resolution multispectral images and high-resolution multispectral images is highly nonlinear, which makes it difficult to express explicitly in a mathematical model.

With the development of AI technology, deep learning-based methods have recently become mainstream and are widely used for the fusion and super-resolution of remote sensing images. The deep learning-based methods usually establish the mapping relationship by learning the mapping between low-resolution images and high-resolution images, using models such as convolutional neural networks (CNNs), generative adversarial networks (GANs), and transformers. For example, CNN or GAN models can convert the 20 m or 60 m bands of Sentinel-2 images into 10 m spatial resolution images (Gargiulo et al., 2019; Lanaras et al., 2018; Romero et al., 2020; Xiong et al., 2020). Low-resolution images can also be upsampled in a simple way, such as by bilinear or bicubic interpolation, without deep learning, but the resulting images are blurred (Lanaras et al., 2018). Deep learning, which implements mapping of low-resolution images to high-resolution images, produces fused images with spatial detail information of high-resolution images and multispectral information of low-resolution images, and it performs well (Dong et al., 2019; Lu et al., 2019; Wang et al., 2023). In particular, the GAN-based approach has shown advantages in making model training more stable, enhancing generalization performance across locations and sensors, and capturing high-dimensional nonlinear features from a large number of sample data (Ma et al., 2019; Xiong et al., 2020). However, this method can provide a better visual effect, but the quality of the reconstructed remote sensing images can deteriorate (Wang et al., 2023). Hence, to further enhance the performance of deep learning-based image fusion and super-resolution, transformers have recently been applied (Cai & Zhang, 2022; Lei et al., 2022).

However, in the previous methods, the spatial resolution of the fusion image does not exceed the maximum limit of the spatial resolution of those satellite images, and in the case of Sentinel-2 images, the spatial resolution is limited to a level much lower than the spatial resolution of the aerial image. Low spatial resolution satellite images do not reflect the geometry of objects on the Earth's surface in detail, and therefore the accuracy of satellite image analysis cannot meet practical requirements. Hence, we propose a method to enhance the spatial resolution of some spectral bands of Sentinel-2 images by a CNN to the maximum resolution of the Sentinel-2 sensors and then fuse them with high spatial resolution aerial images by Gram-Schmidt orthogonalization. The effectiveness of the proposed method was verified by experiments on the generation of fusion images showing vegetation coverage in urban areas (Wang et al., 2016).

The main objective of this study is to explore a practical method for obtaining useful information on the environment from remote sensing images. This paper is organized as follows. First, we describe the method of fusion and super-resolution of remote sensing images by deep learning and Gram-Schmidt orthogonalization. Next, we discuss the experimental and performance evaluation results regarding vegetation coverage in urban areas. Finally, the conclusions are presented.

METHOD

Characteristics of Sentinel-2 Images

We used Sentinel-2 multispectral images and aerial images. The Sentinel-2 images consist of images of several spectral bands with spatial resolution of 10 m, 20 m, and 60 m (Gargiulo et al., 2019; Lanaras et al., 2018). Table 1 shows the Sentinel-2 bands.

Table 1. Sentinel-2 bands

Bands	B1	B2	B3	B4	B5	B6	B7	B8	B8a	B9	B10	B11	B12
Center wavelength [nm]	443	490	560	665	705	740	783	842	865	945	1380	1610	2190
Bandwidth [nm]	20	65	35	30	15	15	20	115	20	20	30	90	180
Spatial resolution [m]	60	10	10	10	20	20	20	10	20	60	60	20	20

Here, the spectral bands with 10 m spatial resolution are the four bands of B2, B3, B4, and B8, the spectral bands with 20 m spatial resolution are the six bands of B5, B6, B7, B8a, B11, and B12, and the spectral bands with 60 m spatial resolution are the three bands of B1, B9, and B10.

Multispectral image data with spatial resolution of 10 m or 20 m are commonly used as the basis for classification and distribution characterization of features on the Earth's surface. For example, they are used for detailed investigations such as land cover, crop distribution, forest species distribution, biomass distribution, and coastal and disaster area mapping. In addition, data from three spectral bands with a spatial resolution of 60 m are used for atmospheric water vapor and aerosol correction, cirrus measurements, etc. (Lanaras et al., 2018).

Basic Procedure of Fusion and Super-Resolution of Remote Sensing Images

In this paper, we will only deal with spectral bands with 10 m spatial resolution and 20 m spatial resolution in Sentinel-2 image data, aiming at the investigation of vegetation coverage in urban areas. The basic procedure of implementing the fusion and super-resolution of multispectral images and aerial images, remote sensing images, by deep learning and Gram-Schmidt orthogonalization, is as follows.

Step 1. Prepare the Sentinel-2 images and aerial images of the same time period of the study area, and use the software ENVI 3.0 for satellite image processing to perform the necessary atmospheric correction, radiometric correction, and topographic correction for the Sentinel-2 image data, and overlap the images of different bands with the same spatial resolution.

Step 2. Using ArcGIS 10.4, a software package for GIS and image processing, perform precise geometric correction using ground control points (GCPs), thereby maximizing the accuracy of the correction and matching the geographical coordinates of the images.

Step 3. Convert the images of the spectral bands with 20 m spatial resolution into images with 10 m spatial resolution by the deep learning model. Then, several bands of images with 10 m spatial resolution can be used as reference images.

Step 4. Select the spectral-band images needed for real problem solving, and then fuse the selected spectral-band images and aerial images by Gram-Schmidt orthogonal transformation and inverse transformation to obtain a multispectral image with spatial resolution of 1 m. For example, if the problem of investigating vegetation coverage in urban areas is to be considered, the spectral bands available for urban green extraction in Sentinel-2 images are the visible bands B2, B3, and B4 (10 m) and the near-infrared band B8 (10 m) and the shortwave-infrared band B11 (20 m), which reflect vegetation cover. The details of steps 3 and 4 are explained separately later.

In general, the fusion of satellite image data is performed between panchromatic and multispectral bands, whereas the Sentinel-2 images have no panchromatic bands. Hence, we used black and white aerial images as a panchromatic substitute. The black and white aerial images used as panchromatic substitutes were produced in the following way. First, the average image of the Sentinel-2 bands converted to 10 m spatial resolution in step 3 is computed. Then, for the previously prepared black and white aerial images, feature matching is performed by histogram matching according to its average image.

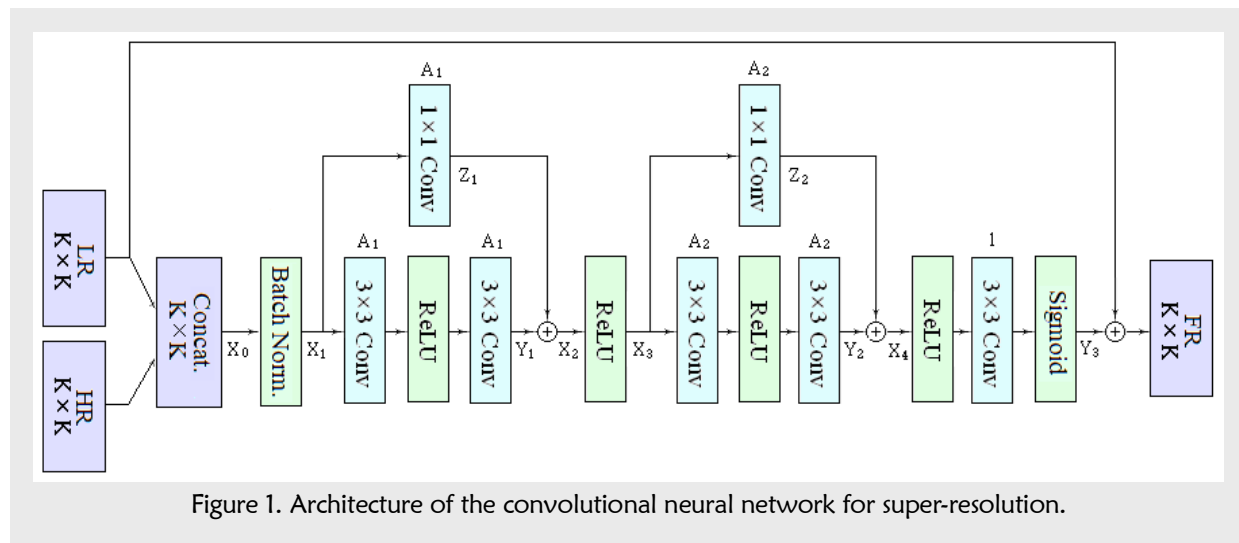
Super-Resolution with Deep Learning

The super-resolution with deep learning is based on the assumption that the spatial detail information reflected in high spatial resolution spectral band images can be efficiently transferred to lower spatial resolution spectral band images, and that the mapping to produce high spatial resolution fusion images can be achieved through deep learning.

Construction of a Deep Learning Model

In this paper, convolutional neural networks with skip connections are used as a deep learning model for the super-resolution of low-resolution satellite images. The skip connection is used to facilitate deep learning for super-resolution and to reduce the information loss that may occur during convolution

operations. The architecture of the convolutional neural network for super-resolution is shown in Figure 1.



The input of the convolutional neural network for super-resolution is the low-resolution image patches LR and high-resolution image patches HR, and the output is the fusion image patches FR of LR and HR, which are equal in size to HR. That is, the convolutional neural network performs image-to-image mapping to produce a fusion image with the same spatial resolution as the HR, containing the information that the LR has. To do this, first, the up-sampling of low-resolution images is performed by bicubic interpolation to make them equal in size. Next, all pixel values of the image are divided by the maximum pixel value so that they lie within the interval $[0, 1]$. During the multiple-stage convolution operation, padding is applied to keep the feature map size unchanged, the convolution step size is 1×1 , and the ReLU (Rectified Linear Unit) function and sigmoid function are applied to avoid loss of the nonlinear characteristics of the neural network.

The flow of the input–output relationship of the convolutional neural network for super-resolution is as follows (Figure 1).

(1) Normalization is performed so that pixels of LR and HR with the same size $K \times K$ have values of interval $[0, 1]$, and then LR and HR are concatenated to form the feature map X_0 of size $K \times K$.

(2) The batch normalization for X_0 is performed to obtain X_1 as the result of the operation.

(3) For X_1 , convolution and ReLU operations are performed, with the size of convolution kernels 3×3 and the number of convolution channels A_1 (e.g., $A_1=32$). Then, without changing the size of the convolution kernel and the number of convolution channels, the convolution operation is performed again. The result of the operation is denoted by Y_1 . Also, for X_1 , convolution operation for skip connection is performed, with the size of convolution kernel 1×1 and the number of convolution channels A_1 . The result of the operation is denoted by Z_1 . Then by adding Z_1 to Y_1 , the result of operation X_2 is obtained.

(4) For X_2 , the ReLU operation is performed to obtain the result of X_3 .

(5) For X_3 , convolution and ReLU operations are performed, with the size of convolution kernels 3×3 and the number of convolution channels A_2 (e.g., $A_2=16$). Then, without changing the size of the convolution kernel and the number of convolution channels, the convolution operation is performed again. The result of the operation is denoted by Y_2 . Also, for X_3 , convolution operation for skip connection is performed, with the size of convolution kernel 1×1 and the number of convolution channels A_2 . The result of the operation is denoted by Z_2 . Then by adding Z_2 to Y_2 , the result of operation X_4 is obtained.

(6) For X_4 , first the ReLU operation is performed, and then the convolution operation is performed. The size of the convolution kernel is also 3×3 and the number of convolution channels is 1. Then, for this convolution result, a Sigmoid function operation with a range of function values $(0, 1)$ is performed. The result of the operation is denoted by Y_3 . Then, by adding LR to Y_3 by skip connection, the fusion image fragment FR as the output of the model is obtained.

For the above input–output relationship, the Batch Normalization (BN) operation is used to stabilize the training process of the neural network, and it is known that in training the convolutional

neural network for super-resolution of satellite images, the first batch normalization before convolution can alleviate the statistical fluctuations of the training data set (Gargiulo et al., 2019). The skip connections in Figure 1 are implemented in steps (3), (5), and (6).

Construction of Sample Dataset for Deep Learning

The construction of the sample dataset for model learning is done as follows. First, the Sentinel-2 image data is tiled into image patches of size $K \times K$. This helps to extract the local spatial texture of the Earth's surface and to understand the details of the objects. In this paper, the size of the HR image patch and FR image patch for high-resolution (10 m resolution) images is 80×80 pixels, and the corresponding LR image patch for low-resolution (20 m resolution) images is 40×40 pixels. When they are used as input to the model, upsampling is performed by bicubic interpolation on LR image patches. Thus, their size is the same as that of the HR image patches.

In this paper, 500 local Sentinel-2 images of 10 locations within the test area for the investigation of vegetation coverage in urban areas were selected as sampling images. Among them, 450 images were used as training and validation data, and the remaining 50 images were used as test data. For the images used as training and validation data, a total of 45,000 image patches were prepared by taking 100 image patches per image; 90% of them (40,500 patches) were used as training data and 10% (4,500 patches) as validation data.

Implementation of Deep Learning for Super-Resolution

As a loss function for learning of the super-resolution model, the mean absolute error (L^1 norm) of the pixel values between the predicted and the reference images was used. The L^1 norm gives better learning results for super-resolution than the mean squared error (L^2 norm) because it is more robust to outliers (Lanaras et al., 2018).

As an algorithm of the deep learning for super-resolution, the Adam optimizer was used (Kingma & Ba 2015). In training, the batch size was set to 64, the initial learning rate was 10^{-4} , and the network was trained for 100 epochs in total. The experiments were run on an Intel(R) Core(TM) i7-7700 CPU at 3.6 GHz (8 threads) and an NVIDIA Titan Xp GPU.

Fusion of Remote Sensing Images by Gram-Schmidt Orthogonalization

In this paper, Gram-Schmidt orthogonal transformation and inverse transformation are used to further improve the performance of satellite image analysis. Thus, high spatial resolution aerial images and low spatial resolution multispectral band images are fused to produce high spatial resolution fusion images.

Principle of Gram-Schmidt Orthogonalization

The Gram-Schmidt orthogonal transformation was proposed by Gram and Schmidt, and the modified Gram-Schmidt method, which is more stable numerically, was analyzed by Björck (1967). In short, the Gram-Schmidt orthogonal transform reconstructs a given set of nonorthogonal, linearly independent vectors into a set of mutually orthogonal normalized vectors.

Now, for two vectors A and B in Euclidean space, let the scalar product of two vectors and the norm of the vector be expressed respectively as follows:

$$(a, b) = (a_1b_1 + a_2b_2 + \Lambda + a_Nb_N), \quad \|a\| = \sqrt{(a, a)}$$

Then, through the Gram-Schmidt orthogonal transformation of a set a_1, a_2, Λ, a_M of linearly independent vectors, the mutually orthogonal vectors b_1, b_2, Λ, b_M , which are not zero vectors, are found as follows:

$$\left. \begin{aligned} b_1 &= a_1 \\ b_2 &= a_2 - \frac{(a_2, b_1)}{(b_1, b_1)} b_1 \\ \Lambda \quad \Lambda \\ b_j &= a_j - \sum_{k=1}^{j-1} \frac{(a_j, b_k)}{(b_k, b_k)} b_k \quad (j = 3, 4, \Lambda, M) \end{aligned} \right\} \quad (1)$$

Now, with respect to Equation (1), the following symbolic expression is introduced.

$$r_{jk} = \frac{(a_j, b_k)}{(b_k, b_k)} \quad (j = 2, 3, \Lambda, M-1, k = 1, 2, \Lambda, j-1) \quad (2)$$

Then the Gram-Schmidt orthogonal transformation to a_1, a_2, Λ, a_M can be expressed from Equations (1) and (2) in the matrix form as follows:

$$(b_1 \ b_2 \ \Lambda \ b_M) = (a_1 \ a_2 \ \Lambda \ a_M) \begin{pmatrix} 0 & r_{21} & r_{31} & \Lambda & r_{M1} \\ 0 & 0 & r_{32} & \Lambda & r_{M2} \\ 0 & 0 & 0 & \Lambda & r_{M3} \\ \Lambda & \Lambda & \Lambda & & \Lambda \\ 0 & 0 & 0 & \Lambda & 0 \end{pmatrix} \quad (3)$$

The following equation holds from (3):

$$(b_1 \ b_2 \ \Lambda \ b_M) \begin{pmatrix} 1 & r_{21} & r_{31} & \Lambda & r_{M1} \\ 0 & 1 & r_{32} & \Lambda & r_{M2} \\ 0 & 0 & 1 & \Lambda & r_{M3} \\ \Lambda & \Lambda & \Lambda & & \Lambda \\ 0 & 0 & 0 & \Lambda & 1 \end{pmatrix} = (a_1 \ a_2 \ \Lambda \ a_M) \quad (4)$$

Now, with respect to Equation (4), the following symbolic expression is introduced.

$$\tilde{R} = \begin{pmatrix} 1 & r_{21} & r_{31} & \Lambda & r_{M1} \\ 0 & 1 & r_{32} & \Lambda & r_{M2} \\ 0 & 0 & 1 & \Lambda & r_{M3} \\ \Lambda & \Lambda & \Lambda & & \Lambda \\ 0 & 0 & 0 & \Lambda & 1 \end{pmatrix}$$

$$A = (a_1 \ a_2 \ \Lambda \ a_M), \quad B = (b_1 \ b_2 \ \Lambda \ b_M),$$

Then, Equation (4) of the Gram-Schmidt orthogonal transform is expressed as follows:

$$A = B \tilde{R} \quad (5)$$

Here $\tilde{R}$ is clearly a nondegenerate matrix. Therefore, it has an inverse, i.e., it is invertible.

The inverse transformation of the Gram-Schmidt orthogonal transform is expressed by the inverse matrix $\tilde{R}^{-1}$ of $\tilde{R}$, from Equation (5) as follows:

$$B = A \tilde{R}^{-1} \quad (6)$$

In this paper, the Gram-Schmidt orthogonal transformation for a_1, a_2, Λ, a_M is denoted by $GST(a_1, a_2, \Lambda, a_M)$ and the inverse transform for b_1, b_2, Λ, b_M is denoted by $GST^{-1}(b_1, b_2, \Lambda, b_M)$.

On the other hand, by normalizing b_1, b_2, Λ, b_M as follows, orthonormal vectors c_1, c_2, Λ, c_M are obtained.

$$c_1 = \frac{b_1}{\|b_1\|}, \quad c_2 = \frac{b_2}{\|b_2\|}, \quad \dots, \quad c_M = \frac{b_M}{\|b_M\|} \quad (7)$$

Application of Gram-Schmidt Orthogonal Transform and Inverse Transformation in Image Fusion

The procedure of performing the fusion of spectral band images and panchromatic images by Gram-Schmidt orthogonal transformation and inverse transformation is as follows.

Step 1. By linear combination of spectral band images a and b , a pseudo-panchromatic image A is generated as follows:

$$PAN' = \sum_{i=1}^M w_i SI_i, \quad \sum_{i=1}^M w_i = 1 \quad (8)$$

Here, w_i is the weight of the i -th spectral band image.

Step 2. For the pseudo-panchromatic image PAN' and the spectral band images $SI_1, SI_2, \dots, SI_M$, the Gram-Schmidt orthogonal transformation $GST(PAN', S_1, \Lambda, S_M)$ is implemented, so that a set

of normalized image vectors $(GS_0, GS_1, \Lambda, GS_M)$ that are mutually orthogonal is produced.

$$(GS_0, GS_1, \Lambda, GS_M) = GST(PAN', S_1, \Lambda, S_M) \quad (9)$$

Step 3. The average image SI_{avg} for the spectral band images $SI_M, SI_1, SI_2, \Lambda, SI_M$ is calculated.

$$SI_{avg} = \frac{1}{M} \sum_{i=1}^M SI_i \quad (10)$$

For the black and white aerial images, the histogram transformation according to the average image SI_{avg} is performed to obtain the panchromatic image PAN . Then, in the mutually orthogonal normalized image vector set $(GS_0, GS_1, \Lambda, GS_M)$, GS_0 is replaced by a normalized panchromatic image vector PAN , so that the normalized image vector set $(PAN, GS_1, \Lambda, GS_M)$ is obtained.

Step 4. For $(PAN, GS_1, \Lambda, GS_M)$, the inverse transformation $GST^{-1}(\cdot)$ is implemented. Thus, the fusion images $(FI_0, FI_1, \Lambda, FI_M)$ as the fusion results of each spectral band image and the panchromatic image are produced.

$$(FI_0, FI_1, \Lambda, FI_M) = GST^{-1}(PAN, GS_1, \Lambda, GS_M) \quad (11)$$

In this paper, only black and white aerial images are used instead of panchromatic images.

RESULTS & DISCUSSION

To verify the effectiveness of the proposed method, we performed experiments on the fusion of aerial images with Sentinel-2 images showing vegetation coverage in urban areas. First, we performed experiments to increase the spatial resolution of some spectral bands with 20 m spatial resolution among Sentinel-2 images by deep learning to 10 m, the maximum resolution of the Sentinel-2 remote sensor. Next, we performed experiments to fuse the aerial images with 1 m spatial resolution and Sentinel-2 multispectral images with 10 m spatial resolution by Gram-Schmidt orthogonalization.

Experimental Results and Discussion on Spatial Resolution Enhancement

The spectral bands available for the investigation of vegetation distribution in urban areas in the Sentinel-2 image are the visible bands B2, B3, and B4, and the near-infrared band B8 and the shortwave-infrared band B11, which reflect vegetation cover. Here, the spatial resolution of the B2, B3, B4, and B8 bands is 10 m, but the spatial resolution of the B11 band is 20 m, so it is necessary to increase the spatial resolution of the B11 band to 10 m. For this B11 band image, deep learning is applied to preserve spectral information and enhance spatial resolution.

As input to the deep learning model, the B2, B3, B4, and B8 bands with 10 m spatial resolution and the B5, B6, B7, B8a, B11 and B12 bands with 20 m spatial resolution were used. As an output of the deep learning model, the image data of bands B5, B6, B7, B8a, B11, and B12 were produced with 10 m spatial resolution. Figure 2 shows the results of spatial resolution enhancement of Sentinel-2 images by the proposed model. Comparing the pseudo-color composite image of B5, B6, and B7 bands with 20 m resolution (c) and the corresponding pseudo-color composite image of B5, B6, and B7 bands with 10 m resolution (e), as shown in Figure 2, we can see that the Sentinel-2 image of these bands is spatially sharper on visual inspection. Furthermore, comparing the pseudo-color composite image of B8a, B11, and B12 bands with 20 m resolution (d) and the corresponding pseudo-color composite image of B8a, B11, and B12 bands with 10 m resolution (f), we can see that the Sentinel-2 image of these bands is spatially sharper.

Quantitative performance comparisons are performed with fusion images using the proposed deep learning model and with two previous methods, bicubic interpolation and area-to-point regression kriging (ATPRK). The evaluation measures are RMSE, CC, SRE, SAM, PSNR, and SSIM.

The root mean square error (RMSE) is defined as follows: the smaller this value, the better the fusion result.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{x}_i - x_i)^2} \quad (12)$$

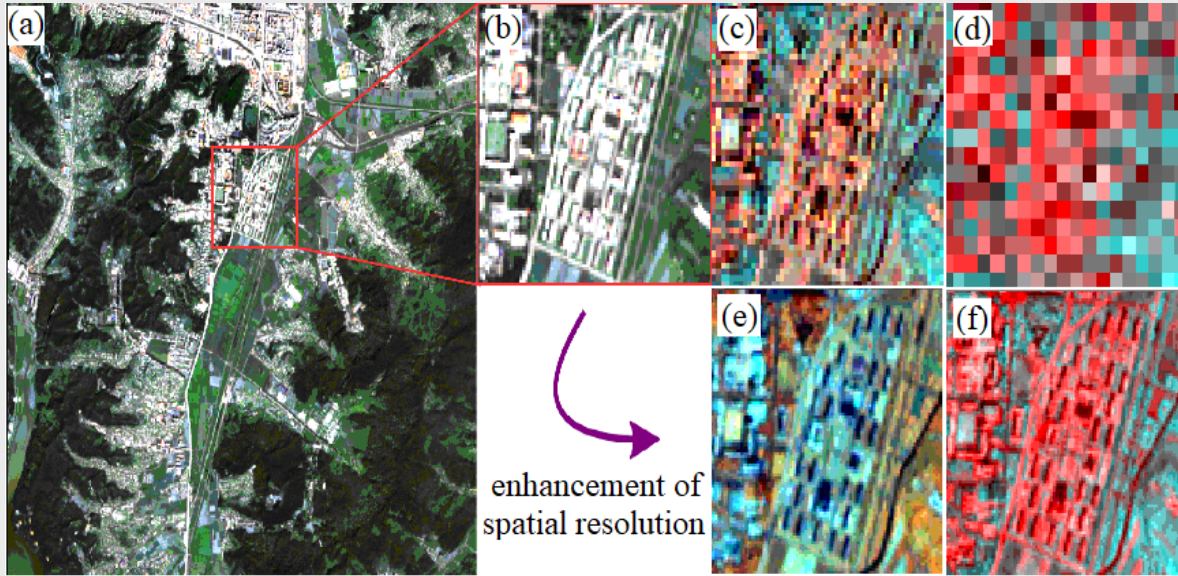


Figure 2. Enhancement of spatial resolution of Sentinel-2 images by the proposed deep learning model: (a) pseudo-color composite images of the 10 m resolution B2, B3, and B4 bands, (b) pseudo-color composite image of the sub-region of (a), (c) pseudo-color composite images of the 20 m resolution B5, B6, and B7 bands, (d) pseudo-color composite image of the B8a, B11, and B12 bands with 20 m resolution, (e) image with spatial resolution enhanced by the proposed model (corresponding to (c)), (f) image with spatial resolution enhanced by the proposed model (corresponding to (d)).

where $\hat{x}_i$ is the i -th pixel value of the fusion image, x_i is the i -th pixel value of the reference image, and n is the total number of pixels.

The correlation coefficient (CC) is defined as follows: the closer the value is to 1, the higher the quality of the fusion image.

$$CC = \frac{\sum_{i=1}^n (\hat{x}_i - \mu_{\hat{x}})(x_i - \mu_x)}{\sqrt{\sum_{i=1}^n (\hat{x}_i - \mu_{\hat{x}})(x_i - \mu_x)}} \quad (13)$$

where $\mu_{\hat{x}}$ is the average of the pixel values of the fusion image and μ_x is the average of the pixel values of the reference image.

The signal-to-reconstruction error ratio (SRE) is used to compare the errors of images with different luminance levels through the errors associated with the average density of the images, the higher the value, the better.

$$SRE = 10 \log_{10} \frac{\mu_x^2}{\|\hat{\mathbf{x}} - \mathbf{x}\|^2 / n} \quad (14)$$

where $\hat{\mathbf{x}}$ is the fusion image vector, $\mathbf{x}$ is the reference image vector, and the value of SRE is calculated in decibels (dB).

The SAM (Spectral Angle Mapper) is defined as follows: the smaller this value, the higher the spectral similarity of the two images.

$$SAM(\mathbf{x}, \hat{\mathbf{x}}) = \arccos\left(\frac{\mathbf{x} \cdot \hat{\mathbf{x}}}{\|\mathbf{x}\| \cdot \|\hat{\mathbf{x}}\|}\right) \quad (15)$$

The PSNR (Peak Signal-to-Noise Ratio) is defined as follows: the larger this value, the higher the quality of the fusion image.

$$PSNR = 10 \log_{10} \frac{1}{MSE(\mathbf{x}, \hat{\mathbf{x}})} \quad (16)$$

$$MSE(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{h \cdot w \cdot c} \sum_{k=1}^c \sum_{i=1}^h \sum_{j=1}^w [\mathbf{x}(i, j, k) - \hat{\mathbf{x}}(i, j, k)]^2 \quad (17)$$

The SSIM (Structural Similarity) is defined as follows: the larger the value, the higher the quality of the fusion image.

$$SSIM = \frac{(2\mu_x\mu_y + C_1)(2\sigma_x\sigma_y + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (18)$$

$$C_1 = (k_1 L)^2, \quad C_2 = (k_2 L)^2$$

where C_1 and C_2 are stabilizing constants representing the change of pixel values (e.g., $k_1 = 0.01$, $k_2 = 0.03$, $L = 1$).

The quantitative performance comparison of the proposed model with the previous methods is shown in Table 2 and Figure 3.

Table 2. Quantitative performance comparison of the proposed model with the previous methods.

model	RMSE	CC	SRE	SAM	PSNR	SSIM
Bicubic	123.4	0.811	25.1	1.25	15.781	0.772
ATPRK	117.8	0.866	26.6	1.67	29.726	0.895
proposed model	43.5	0.941	37.5	0.78	38.544	0.981

The quality of image fusion by the proposed model is considerably better than that of the previous methods, bicubic interpolation and ATPRK, as shown in Table 2 and Figure 3. The quantitative comparison results by RMSE, CC, SRE, SAM, PSNR and SSIM in Table 2 show that the proposed model improves on both previous methods in all six evaluation measures. Figure 3 also qualitatively shows that (d), (h) and (i) by the proposed model are better quality fusion images than (b), (f), (j) by bicubic interpolation or (c), (g) by ATPRK.

Experimental Results and Discussion on Remote Sensing Image Fusion

For the purpose of investigation of vegetation coverage in urban areas, the fusion of the spectral images of bands B2, B3, B4, B8 and B11 with 10 m spatial resolution (B11 enhanced by deep learning) and the aerial images with 1 m spatial resolution was performed. The application of Gram-Schmidt orthogonal transformation and inverse transformation to the image fusion produced the fusion images with 1 m spatial resolution. Some experimental results on remote sensing image fusion are discussed as follows.

First, Figure 4 shows the pseudo-color composite images of bands B2, B3 and B4 with 10 m spatial resolution for the test area, aerial images with 1 m spatial resolution, and pseudo-color composite images of bands B2, B3 and B4 as the result of image fusion. Comparing the Sentinel-2 image with the corresponding aerial image and the fusion image, as shown in Figure 4, it can be seen from visual analysis that the fusion image showing the vegetation distribution in urban areas was as sharp as the aerial image with 1 m spatial resolution.

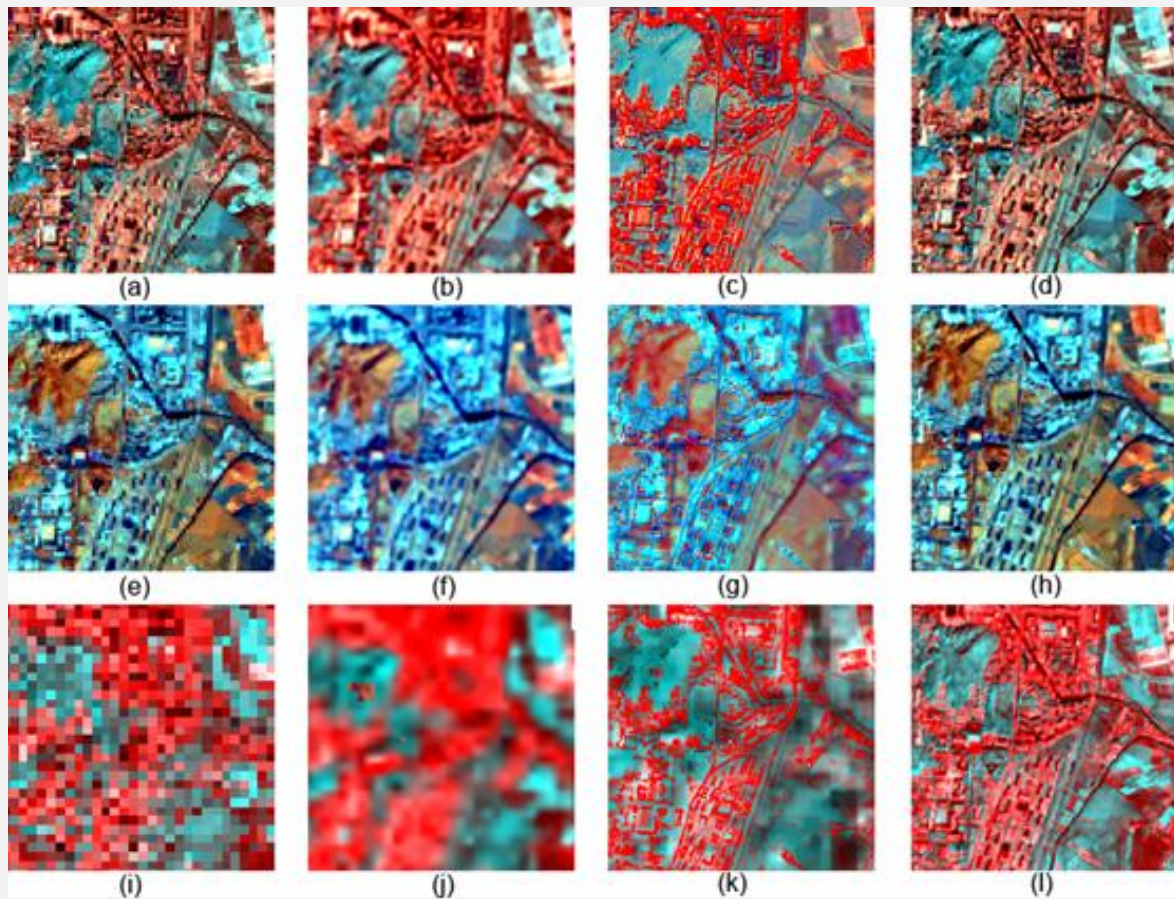


Figure 3. Comparison of Sentinel-2 fusion images using different models: (a, e, i) original image of Sentinel-2, (b, f, j) Bicubic, (c, g, k), ATPRK, (d, h, l), (a) pseudo-color composite image of bands B5, B6 and B7, (e) pseudo-color composite image of bands B8A, B11 and B12, (i) pseudo-color composite image of bands B1, B9 and B10.

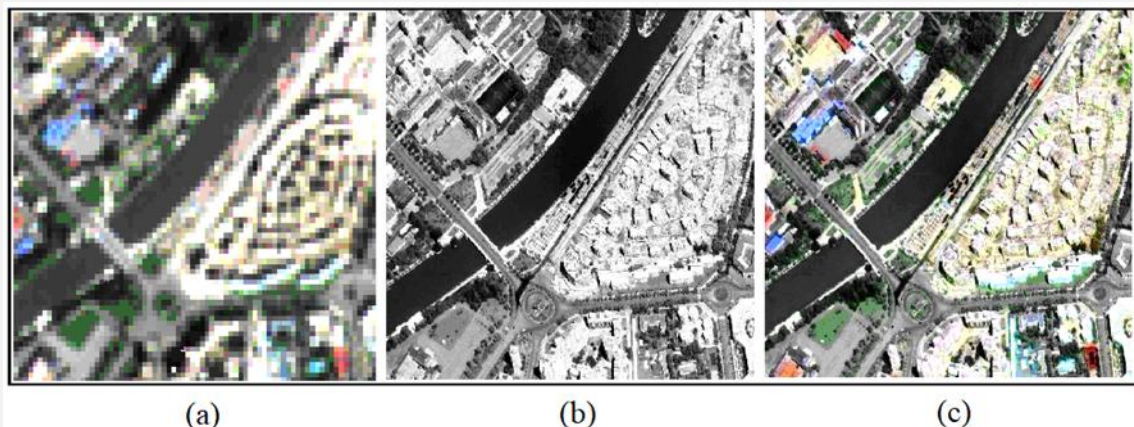


Figure 4. Comparison of Sentinel-2 image, aerial image and fused image: (a) Sentinel-2 image (pseudo-color composite image of bands B2, B3 and B4 with 10 m spatial resolution), (b) aerial image with 1 m spatial resolution, and (c) fused image (pseudo-color composite images of bands B2, B3 and B4 as the image fusion results).

Next, Figure 5 shows a pseudo-color composite image of the B2, B3 and B11 bands as the result of remote sensing image fusion by the proposed method for the test area.



Figure 5. A pseudo-color composite image of the B2, B3 and B11 bands as the result of remote sensing image fusion by the proposed method.

Here, the B11-band image is originally at 20 m spatial resolution, which was enhanced to 10 m spatial resolution through deep learning. As shown in Figure 5, the remote sensing image fusion by the proposed method can be used to show the vegetation distribution of urban areas well and produce a sharp fusion image comparable to a 1 m aerial image.

Next, the spectral characteristic curves of the original low-resolution and fused high-resolution images for urban green space and road area are shown in Figure 6 and Figure 7, respectively.

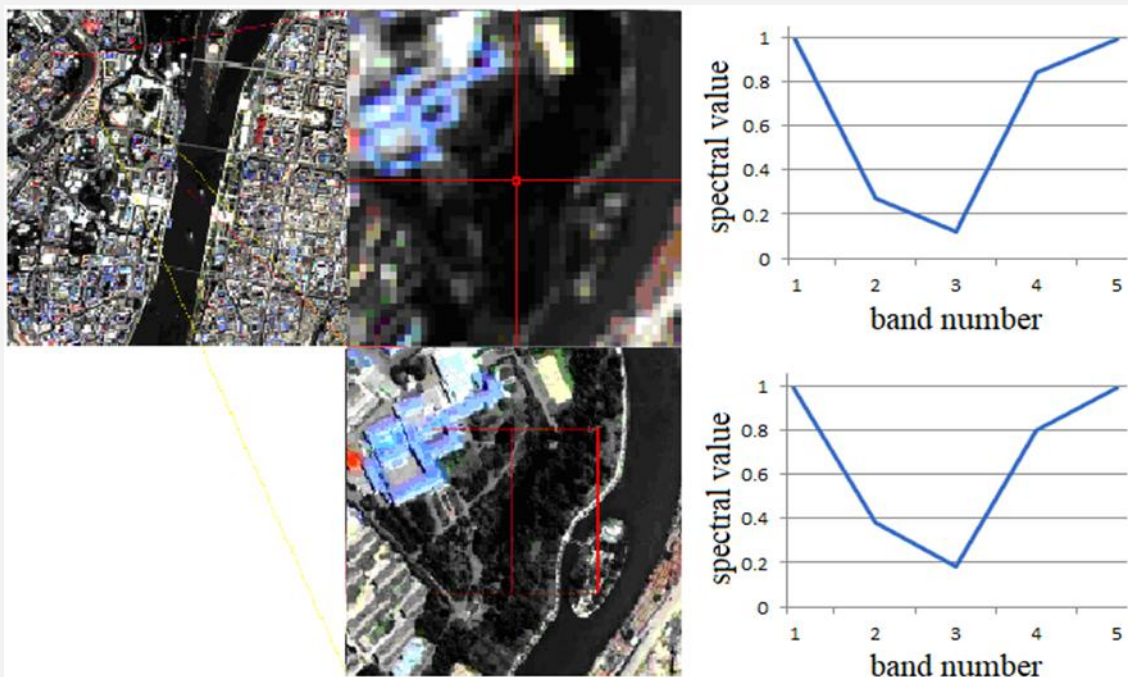


Figure 6. Comparison of spatial resolution and characteristic curves of the original low-resolution and fused high-resolution images for urban green space.

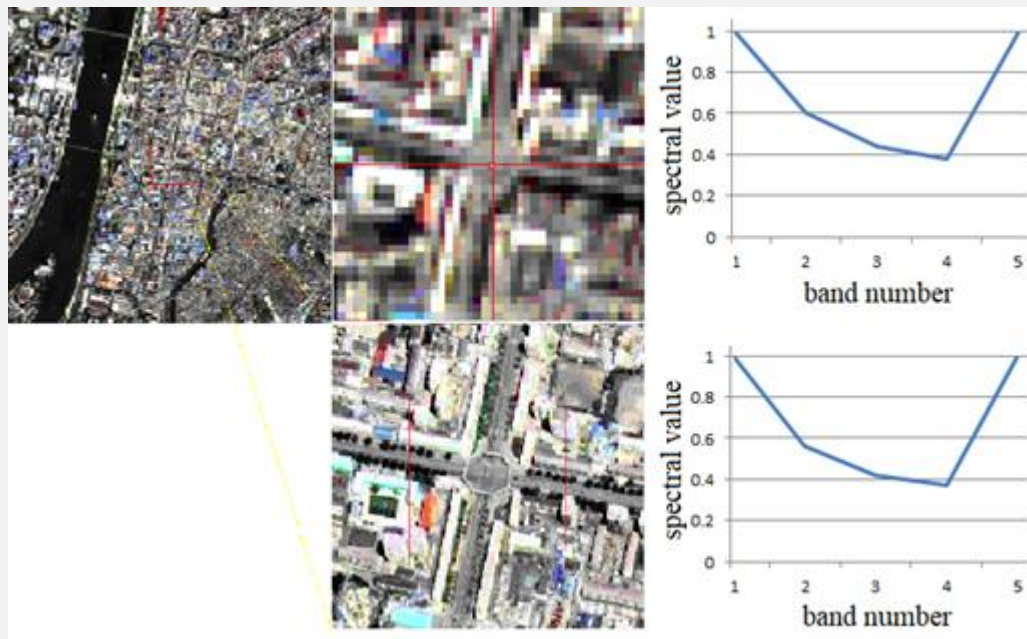


Figure 7. Comparison of spatial resolution and characteristic curves of the original low-resolution and fused high-resolution images for urban road area.

As shown in Figure 6 and Figure 7, it can be seen that the original low-resolution images and fused high-resolution images for urban green space and road areas differ greatly in spatial resolution, but the shape of the spectral characteristic curve is similar. This shows that the multispectral images produced by the proposed remote sensing image fusion method preserve the inherent meaning of the corresponding features in terms of spectral characteristics.

It should be emphasized that our model using Gram-Schmidt orthogonal transformation and inversion theoretically assumes linear independence of spectral bands, but in practice it is possible to obtain visually good fusion results, although the spectral bands are somewhat dependent. As shown in Figures 4 to 7, the B2, B3, B4, and B11 bands of the Sentinel-2 image are somewhat dependent, but visually good fusion results are obtained by the proposed model.

Next, Tables 3 and 4 show the quantitative comparison of the original image (Sentinel-2 image of the study area) and the fusion image. Since the Sentinel-2 data had the highest spatial resolution of 10 m and no spatial resolution of 1 m, quantitative estimation of fusion results was performed by reducing the spatial resolution of the fusion image to 10 m to estimate the RMSE and CC with the original image.

Table 3. Global spectral mean values of the original and fused images.

	B2	B3	B4	B8	B11	mean
original image	490	560	665	842	1610	833.4
fusion image	508	536	673	820	1508	809.0

Table 4. Quantitative comparison of fusion and original images.

	B2	B3	B4	B8	B11	mean
RMSE	43.2	45.2	43.9	48.2	83.3	52.76
CC	0.93	0.91	0.93	0.9	0.81	0.896

As shown in Table 3, the spectral mean of the fusion image was 0.971 times that of the original image (a 2.9% decrease). Also, as shown in Table 4, the mean RMSE and CC between the fusion image and the original image over the five bands are 52.76 and 0.896, respectively.

These results suggest that the multispectral fusion image produced by the proposed model may be useful for practical applications.

CONCLUSION

In this paper, we proposed a method to realize fusion and super-resolution of remote sensing images, Sentinel-2 images and aerial images, by deep learning and Gram-Schmidt orthogonalization, to further enhance the performance of remote sensing image analysis of the Earth's environment. The main points of our method are as follows. First, by deep learning, a model is constructed in which the spatial detail information reflected in high spatial resolution spectral band images can be efficiently transferred to low spatial resolution spectral band images, while the spatial resolution of some spectral band images of Sentinel-2 is enhanced to the maximum resolution of the remote sensor of Sentinel-2. Using this model, the spectral band images with a spatial resolution of 20 m are enhanced to 10 m. Next, the process of fusing the 10 m spatial resolution-matched Sentinel-2 images with 1 m spatial resolution aerial images is performed using Gram-Schmidt orthogonal transformation and inverse transformation. Thus, the image fusion, which reflects the information contained in the low-resolution multispectral image, is realized. To verify the effectiveness of the proposed method, we performed experiments on the fusion of black and white aerial images and Sentinel-2 images with different spatial resolution and spectral bands showing vegetation coverage in urban areas. Experimental results show that the proposed method can produce visually fine fusion images while preserving spectral information well. The significance of this study is that it allows us to visually better extract the significant information on the environment from remote sensing images such as Sentinel-2 images and black and white aerial images, thus further enhancing the efficiency of satellite information analysis.

REFERENCES

- Aiazzi, B., Baronti, S., & Selva, M. (2007). Improving component substitution pansharpening through multivariate regression of MS + Pan data. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10), 3230-3239. <https://doi.org/10.1109/TGRS.2007.901007>
- Björck, Å. (1967). Solving linear least squares problems by Gram-Schmidt orthogonalization. *BIT*, 7(1), 1-21. <https://doi.org/10.1007/BF01934122>
- Cai, D., & Zhang, P. (2022). T³SR: Texture transfer transformer for remote sensing image superresolution. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 7346-7358. <https://doi.org/10.1109/jstars.2022.3198557>
- Dong, X., Xi, Z., Sun, X., & Gao, L. (2019). Transferred multi-perception attention networks for remote sensing image super-resolution. *Remote Sensing*, 11(23), Article 2857. <https://doi.org/10.3390/rs11232857>
- Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., ... & Bargellini, P. (2012). Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote sensing of Environment*, 120, 25-36. <https://doi.org/10.1016/j.rse.2011.11.026>
- Gargiulo, M., Mazza, A., Gaetano, R., Ruello, G., & Scarpa, G. (2019). Fast super-resolution of 20 m Sentinel-2 bands using convolutional neural networks. *Remote Sensing*, 11(22), Article 2635. <https://doi.org/10.3390/rs11222635>
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. <https://doi.org/10.48550/arXiv.1412.6980>
- Lanaras, C., Bioucas-Dias, J., Galliani, S., Baltsavias, E., & Schindler, K. (2018). Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 146, 305-319. <https://doi.org/10.1016/j.isprsjprs.2018.09.018>
- Lei, S., Shi, Z., & Mo, W. (2022). Transformer-based multistage enhancement for remote sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 60, Article 5615611. <https://doi.org/10.1109/tgrs.2021.3136190>
- Lu, T., Wang, J., Zhang, Y., Wang, Z., & Jiang, J. (2019). Satellite image super-resolution via multi-scale residual deep neural network. *Remote Sensing*, 11(13), Article 1588. <https://doi.org/10.3390/rs11131588>
- Ma, W., Pan, Z., Yuan, F., & Lei, B. (2019). Super-resolution of remote sensing images via a dense residual generative adversarial network. *Remote Sensing*, 11(21), Article 2578. <https://doi.org/10.3390/rs11212578>
- Romero, L. S., Marcello, J., & Vilaplana, V. (2020). Super-resolution of Sentinel-2 imagery using generative adversarial networks. *Remote Sensing*, 12(15), Article 2424. <https://doi.org/10.3390/rs12152424>
- Vivone, G., Alparone, L., Chanussot, J., Dalla Mura, M., Garzelli, A., Licciardi, G. A., ... & Wald, L. (2014). A critical comparison among pansharpening algorithms. *IEEE Transactions on Geoscience and Remote Sensing*, 53(5), 2565-2586. <https://doi.org/10.1109/TGRS.2014.2361734>
- Wang, Y., Shao, Z., Lu, T., Huang, X., Wang, J., Chen, X., Huang, H., & Zuo, X. (2023). Remote sensing image super-resolution via multi-scale texture transfer network. *Remote Sensing*, 15(23), Article 5503. <https://doi.org/10.3390/rs15235503>
- Wang, Q., Shi, W., Li, Z., & Atkinson, P. M. (2016). Fusion of Sentinel-2 images. *Remote sensing of environment*, 187, 241-252. <https://doi.org/10.1016/j.rse.2016.10.030>
- Xiong, Y., Guo, S., Chen, J., Deng, X., Sun, L., Zheng, X., & Xu, W. (2020). Improved SRGAN for remote sensing image



Copyright (c) 2026 by the authors. This work is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/).